EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Doktoretza-tesia

Txosten klinikoak euskararen eta gazteleraren artean itzultzen laguntzeko corpusaren bilketa eta itzultzaile automatikoaren garapena

Xabier Soto García

2021

EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Txosten klinikoak euskararen eta gazteleraren artean itzultzen laguntzeko corpusaren bilketa eta itzultzaile automatikoaren garapena

Xabier Soto Garciak Gorka Labaka Intxauspe eta Maite Oronoz Anchordoquiren zuzendaritzapean eginiko tesi-txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2021eko azaroa.

Gaien aurkibidea

Taulen zerrenda

Irudien zerrenda

1. KAPITULUA

Sarrera

Gaur egun, itzultzaile automatikoak gure eguneroko bizitzaren parte bilakatu dira, izan modu kontzientean zalantzaren bat argitzeko online tresnetara jotzen dugunean (besteak beste, elia.eus edo batua.eus), zein modu inkontzientean paperean edota pantailaren aurrean automatikoki itzulitako testuren bat irakurtzen dugunean.

Badakigu, ordea, itzultzeko tresnen arabera, hauek elikatzeko erabili diren corpusen arabera, eta jatorri- eta helburu- hizkuntzen arabera, itzulpen automatikoaren kalitatea asko alda daitekeela. Esan beharrik ez dago, publikatu baino lehen itzulpen automatikoa pertsona batek ikuskatu duen ala ez, pertsona hori itzultzaile profesionala den ala ez, edota testuaren domeinuaren ezagutza duen ala ez, faktore erabakigarriak izango direla sortutako itzulpenaren kalitatea ebaluatzerakoan.

Ikasketa automatikoan oinarritutako itzultzaile automatikoei dagokienez, eskuragarri dauden itzultzaile automatikoak edozein motatako testuak itzultzeko garatuak izan dira, ikasketa-prozesuan gehienbat publikoak diren testu elebidunak erabiliz. Hortaz, sarean ohikoak diren testuak (adib.: albisteak) modu egokian itzultzeko gai diren bezala, pribatuak diren txosten klinikoak itzultzerakoan arazoak izango dituzte, besteak beste, testu hauetan erabiltzen den terminologia ez dutelako behar bezala ezagutzen.

Orokorrean, gaur egun gehien erabiltzen den teknika corpusetan oinarritutako itzulpen automatikoa izanda, zenbat eta esaldi pare elebidun gehiago izan itzuli nahi den hizkuntza pare eta domeinurako, hainbat eta handiagoa izango da garatutako itzultzaile automatikoaren kalitatea.

Horretaz gain, behin itzultzaile automatikoaren hizkuntza parea definitua izanik, itzulpenaren zentzuak ere eragina izango du itzulpenen kalitatean. Izan ere, helburu-hizkuntzako eta domeinuko esaldi elebakar kopurua handitzearekin batera, garatutako itzultzaile automatikoaren kalitatea ere handituko da.

Gure inguruan, euskaraz idatzitako testuak bizilagun ditugun gaztelerazko eta frantsesezko testuak baino gutxiago izanda, itzulpenaren kalitatea beti izango da hobea euskaratik baliabide gehiago dituen beste hizkuntza batera egitean, alderantzizko zentzuan egiten denean baino.

Hortaz, bai itzulpen automatikoen maila hobea izateko, baita euskarazko edukien sorkuntza sustatzeko, euskara jatorri- edo helburu- hizkuntza duen itzultzaile automatiko bat garatzerakoan lehentasunezkoa izan beharko luke euskaratik eta ez euskarara itzultzeak.

Gure kasuan, Osakidetzako langile publikoek osasun-txostenak euskaraz idatz ditzaten lagungarria izan daitekeen tresna bat garatzerakoan, irizpide hori jarraitu dugu, diseinatutako sistemek lehentasunezko helburu hori zutelarik. Zentzu honetan, domeinu klinikoaren kasuan ere Osakidetzan bertan gaztelerazko testu kopuru handia izatea lagungarria da, aurrerago azalduko dugun *back-translation* edo atzeranzko itzulpena deritzon teknikari esker.

Edonola ere, domeinu klinikoko testu batean itzulpenean egindako akats batek izan litzakeen ondorioak kritikoak izanik, garatutako itzultzailearen kalitatea edozein izanda ere, bere irteera ikuskatua izan dadin gomendatzen dugu, mediku edo paziente gazteleradun elebakar batek itzulpen automatikoa irakurri baino lehen honen esanahia jatorrizko euskarazko testuaren baliokidea dela bermatuz.

Helburu horrekin, itzultzaile automatikoak diseinatzerakoan itzulpenen zehaztasuna (jatorrizko ingelesez, *adequacy*) ahalik eta altuena izan dadin saiatu gara, terminologia klinikoaren ezagutza eta honen itzulpenen egokitasuna bilatuz. Halaber, kalitate handiena duten garatutako azken sistemak ebaluatzerakoan itzulpen hauek zuzentzeko edo post-editatzeko beharrezkoa den denbora neurtu dugu, Osakidetzako langileentzat erabilgarriena izan daitekeen sistema modu egokian aukeratu ahal izateko.

1.1 Motibazioa

Tesi hau, eta bertan deskribatzen den domeinu klinikoko itzultzaile automatikoa, euskararen normalkuntza prozesuan ekarpen bat izateko asmoarekin

garatuak izan dira.

Dakigun bezala, euskararen ezagutza maila ez dator bat bere erabilera mailarekin, zenbait eremu espezializatuetan desoreka hau nabarmen handiagoa delarik, besteak beste, medikuntzan. Izan ere, gure eguneroko bizitzako hainbat arlotan gertatzen den bezala, Euskal Herriko mediku eta erizainek hegoaldean zein iparraldean gailentzen diren gaztelera edo frantsesa erabiltzen dituzte beraien artean eta pazienteekin idatziz komunikatzeko, nahiz eta elkarrizketan parte hartzen dutenak euskal hiztunak izan.

Tesi hau Euskal Autonomia Erkidegoan kokatzen da, eta hemen osasun-txostenetan gaztelera gailentzearen arrazoietako bat Osakidetzak pazienteen informazioa biltzeko sistema zentralizatua izatea da. Hau horrela izanda, osasun-langile batzuek euskara ez jakiteak erizain eta mediku euskaldunek txosten klinikoak euskaraz idaztea oztopatzen du, pazientearen segurtasuna ez litzatekeelako bermatua egongo.

Honek, ordea, paziente zein osasun-langile euskaldunen eskubide linguistikoak urratzen ditu, ezin baitute informazioa jaso edota sortu beraien lehen hizkuntzan. Bestalde, jakina da arreta jasotzen den hizkuntzak eragin zuzena duela jasotako arretaren kalitatean, beraz, osasun-langileek txosten klinikoak euskaraz idazteak eremu linguistikoaz haratagoko onurak ekarriko lituzke, paziente euskaldunen osasun-arreta bera ere hobetuz.

Zentzu honetan, Osakidetzarekin elkarlanean aurrera eramandako Itzulbide proiektuarekin eta honen barnean garatutako itzultzaile automatikoarekin, osasun-langileek txosten klinikoak euskaraz idazteko lagungarri izan daitekeen tresna bat garatu dugu. Are gehiago, Itzulbide proiektuaren barnean itzultzaile automatikoa garatu ahal izateko jasotako corpus elebiduna bera ere aurrerapauso bat izan da, betidanik gazteleraz idatzi duten mediku eta erizainak euskaraz idazten hasi direlako, testu klinikoak euskaraz idazteko ohitura sortuz, eta beraien artean euskarazko terminologia klinikoaren definizioan sakonduz.

1.2 Lanaren kokapena

Tesi hau Hizkuntzaren Prozesamendua (HP) bezala ezagutzen den arlo akademikoan kokatzen da, Euskal Herriko Unibertsitatean Ixa taldea izanik arlo honetako erreferentzietako bat. Ixa ikerketa-taldea duela hiru hamarkada sortu zenetik aitzindaria izan da HPko hainbat azpiarlotan, bai euskarazko eta euskararako hizkuntza-tresnak sortzen, baita tresna hauek garatzeko

beharrezkoak diren corpusak bildu eta mantentzen. Nazioarteari begira, Europa eta mundu mailako hainbat ikerketa-proiektuetan parte hartzen du, eta zentzu horretan, beste hizkuntza batzuetarako ere baliagarriak diren ikerketak aurrera eraman ditu.

HPk barnebiltzen dituen azpiarlo desberdinetatik, tesi hau Itzulpen Automatikoaren (IAren) arloan kokatzen da. Azken urteetan IA teknika desberdinak garatu ahala, euskara jatorri edota helburu duten tresna desberdinak sortu dira, hala nola, Erregeletan Oinarritutako Itzulpen Automatikoa (EOIA) erabiltzen duen *Matxin* (Mayor, 2007), Itzulpen Automatiko Estatistikoa (IAE) darabilen *EUSMT* (Labaka, 2010), azken bi teknikak konbinatzen dituen SMatxinT (Labaka *et al.*, 2014), eta gaur egun artearen egoera den Itzulpen Automatiko Neuronala (IAN) oinarri hartuta garatutako *MODELA* (Etchegoyhen *et al.*, 2018).

Tesi honetan garatu den itzultzaile automatikoa IANean oinarritua dago, baina garatutako sistema batzuetan EOIA eta IAE sistemak ere erabili dira atzeranzko itzulpena egiteko. Horretaz gain, gure sistema elikatzeko erabili den terminologia klinikoetako bat, SNOMED CT (jatorrizko ingelesez, *Systematized Nomenclature of Medicine – Clinical Terms*) (IHTSDO, 2014), modu automatikoan euskaratua izan da (Perez-de Viñaspre, 2017), hau egiteko hiztegiak, eskuz definitutako erregelak eta EOIA bera ere erabili direlarik.

IAtik haratago, tesi hau medikuntza-domeinuko HPan ere kokatzen da, esparru honetan ere Ixa taldeak esperientzia luze eta zabala duelarik. Zentzu honetan, hainbat proiektu dira Osakidetzarekin batera garatu ahal izan direnak, beraien datu pribatuak erabiliz haien beharretara moldatu diren tresnak garatuz. Tesi hau beste ekarpen bat da eremu honetan, eta aurreko paragrafoan aipatutako lanetan oinarritzen da IA domeinu klinikora egokitzeko.

Proiektuei dagokienez, tesi-lan hau medikuntzarekin lotutako PROSAMED (jatorrizko gazteleraz, *PROcesamiento Semántico textual Avanzado para la detección de dignósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos*) ikerkuntza-proiektuaren barruan kokatzen da, eta tesia aurreratu ahala IAn zentratutako DOMINO (*Traducción Automática Neuronal, en DOMInio, NO supervisada*) eta Tando (*Métodos y Sistemas de Traducción Automática Neuronal Coherente*) proiektuen barruan ere kokatu da.

1.3 Testuingurua

Gaur egun, Euskal Herrian bizi garen helduen % 28,4k erabiltzen dugu euskara (751.500 hizlari aktibo), eta % 16,4 gehiago euskara ulertzeko gai dira (434.000 entzule pasibo), guztira 1.185.500 pertsona izanda (% 44,8). Hizlari aktiboetatik, 700.300 pertsona espainiar estatuan kokatzen diren Araba, Bizkaia, Gipuzkoa eta Nafarroa garaian bizi dira; gainontzeko 51.200ak frantziar estatuan bizi direlarik, Lapurdi, Nafarroa Beherea eta Zuberoan.[1]

Azken hamarkadetan hezkuntzan eman den euskalduntze-prozesuari esker, portzentaia hauek altuagoak dira 16 urtetik azpikoak ere kontuan hartzen baditugu. Horrela, tesi honen kokapen den Euskal Autonomia Erkidegoan, 2 urtetik gorako 895.942 hizlari aktibo eta 391.897 entzule pasibo zenbatzen dira.[2] Tesi-lan hau hasteko momentuan, Osakidetzako langile finkoen % 46,60 elebiduna zen, eta aldi-baterako langileen artean portzentaia hau % 56,29ra igotzen zen.

Euskara hizkuntz ofiziala da Euskal Autonomia Erkidegoan eta Nafarroa garaiko iparraldean, baina euskararen estandarizazio prozesuaren abiatzea eta Osakidetzaren moduko instituzioen sorrera duela 4-5 hamarkada besterik ez ziren gertatu. Gauzak horrela, oraindik ere badaude eremuak non gazteleraren erabilera guztiz gailentzen den euskararen erabileraren aurrean.

Osasunaren arloan hainbat arrazoi daude horretarako: mediku eta erizain guztiek euskara ez ulertzea, duela urte gutxira arte ikasketak euskaraz egin ahal ez izatea, edota euskarazko terminologia klinikoa estandarizazio prozesuan egotea.

Edonola ere, nahiz eta idatziz gaztelera erabili, Osakidetzako langile eta paziente euskaldunen arteko elkarrizketak euskaraz egin ahal izan dira, eta geroz eta gehiago dira txosten klinikoak euskaraz idatzi nahiko lituzketen osasun-langileak.

Azken hamarkadan hainbat aurrerapauso eman dira osasun-txostenak euskaraz idatz daitezen laguntzeko, besteak beste, Donostia Unibertsitate Ospitaleak argitaratutako euskarazko txosten klinikoen ereduen bilduma (Joanes Etxeberri Saria V. Edizioa, 2014), zeina medikuntzako ikasleek modu egokian idazteko helburuarekin argitaratu zen; edota nazioarteko erreferen-

[1] 2016ko VI. Inkesta Soziolinguistikoa: `https://www.irekia.euskadi.eus/uploads/attachments/9954/VI_INK_SOZLG-EH_eus.pdf?1499236557`

[2] Eustat: `https://eu.eustat.eus/bankupx/pxweb/eu/euskara/-/PX_3671_ne02.px/table/tableViewLayout1/`

tzia den GNS-10 Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioan jasotako deskripzioen euskaratzea,[3] azken urteetan mundu mailako IAko erreferentziazko konferentzia den WMTn (jatorrizko ingelesez, *Conference on Machine Translation*) ingelesaren eta euskararen artean biomedikuntzaren arloko testuak itzultzeko atazan parte-hartzaileentzako eskuragarri utzi dena (Bawden *et al.*, 2020).

Azken urteetan sare neuronal artifizialak erabilita HPn eta IAn emandako aurrerapenak kontuan hartuta, tesi honen hasiera momentu egokia kontsideratu zen osasun-langileek euskaraz idazteko lagungarri izan zitekeen itzultzaile automatiko bat garatzeko, hau entrenatu eta ebaluatzeko beharrezkoa den corpus elebiduna osasun-langileek beraiek bilduko zutelarik.

1.4 Helburuak

Tesi honen helburu nagusia txosten klinikoak euskaratik gaztelerara modu egokian itzultzeko gai den itzultzaile automatiko bat garatzea da. Bide horretan, hainbat dira igaro beharreko mugarriak; besteak beste, aipatu dugun atzeranzko itzulpena aplikatu ahal izateko, beharrezkoa da gazteleratik euskarara txosten klinikoak itzultzeko tresna bat garatzea, berau ere tesiaren helburuetako bat izanik.

Itzulpen-zentzua edozein izanda ere, corpusetan oinarritutako teknika aurreratuenak erabili ahal izateko, gure atazara egokitzen diren corpus elebidun eta elebakarrak beharrezkoak dira. Zentzu honetan, tesiaren helburu garrantzitsuenetako bat txosten klinikoak euskararen eta gazteleraren artean itzultzeko corpus egokiak biltzea da. Honekin batera, terminologia klinikoa modu egokian itzuli ahal izateko, euskaraz (eta gazteleraz) eskuragarri dauden terminologikoa klinikoak bildu eta itzulpen-sistemetan modu egokian integratzeko moduak bilatzea da beste helburuetako bat.

Corpusak biltzearekin batera, artearen egoera diren teknikak konparatzea da gure helburuetako bat, horien artean IAN tresna desberdinak, IAE sistemak eta corpusik behar ez duen EOIA ere kontuan hartu direlarik. IANaren barruan, azken urteetan aurkeztu diren aurrerapen desberdinak probatzea da beste helburuetako bat, tartean arkitektura desberdinak eta hitzen segmentazioa egiteko teknika desberdinak egonik.

Zeharkako helburuetako bat IAN sistema desberdinak linguistikoki azter-

[3] `https://drive.google.com/drive/folders/1gUQHoutvYIXGGPVTBbBF3qlHHhX9qbrO`

tzea da, atzeranzko itzulpena egiteko sistemak ebaluatzerakoan sistemen IA metrikak kalkulatzeaz gain, sortutako corpusen aniztasun lexikala (AL) ere neurtuz. Analisi honen barruan, bildutako corpusean aurki daitezkeen genero alborapenak ere neurtzea helburuetako bat da, euskarazko generorik gabeko 'erizain' eta 'mediku' terminoak gazteleraz zein formatan idazten diren aztertuz.

Tesi honen beste helburuetako bat aurkeztutako aurrerapenak nazioarteko testuinguruan kokatzea da, gure atazarako baliagarriak diren irtenbideak besteentzako ere lagungarriak izan daitezkeela frogatzeko asmoz. Helburu horrekin, garatutako teknikak beste hizkuntza pareetan (alemana - ingelesa, ingelesa - gaztelera eta gaztelera - ingelesa) eta antzeko domeinu batean (biomedikuntza) ere aplikatu dira.

Azkenik, itzultzaileen garapenak ingurugiroan izan lezakeen eragina neurtzea da kontuan hartutako beste helburuetako bat. Horrela, esperimentu batzuetan IAN sistemak entrenatu ahal izateko beharrezkoak diren GPUen (jatorrizko ingelesez, *Graphics Processing Unit*) kontsumo energetikoa neurtu dugu, etorkizunean hau murrizteko helburuarekin.

1.5 Arriskuak eta mugak

Edozein aurrerapen teknologikok bezala, IAk ere bere arriskuak ditu; eta gure kasuan, garatuko dugun sistema inplementatu egingo dela aurretik badakigunez, beharrezkoa da honek jendartean izan ditzakeen ondorioei buruz aldez aurretik hausnartzea. Zentzu honetan, azpiatal honen hasieran gardentasunez mahaigaineratu nahi ditugu itzultzaile automatikoaren inguruan besteek plazaratutako kezkak, horiei aurre egiten dieten iritziak ere aipatuz.

Osasun arlotik haratago, IAn egondako aurrerapenek datozen urteetan euskarazko edukien sorkuntzan izan dezaketen eraginak kezka sortzen du, euskarak inguruko hizkuntzek baino baliabide gutxiago izanik, euskaldunek beste hizkuntzetan sortu eta automatikoki euskaratutako edukiak lehenetsi baititzakete, zuzenean euskaraz sortutako edukiak (albisteak, ikus-entzunezkoak, etab.) alboratuz. Kezka hauek Joxe Rojasek sarean plazaratu zituen modu labur eta argian duela gutxi,[4] eta Igor Leturiak ikuspuntu tekniko batetik erantzun zion, bestelako hizketa-teknologiak ere aipatuz.[5] Tesi honen

[4] `https://www.sarean.eus/traduttore-traditore/`

[5] `https://www.sarean.eus/hizkuntza-eta-hizketa-teknologiak-arriskuak-aukerak-ekidinezinak-beharrezkoak/`

egileak azken artikuluan adierazitakoak bere egiten ditu, euskararako IA ez garatzearen arriskuak handiagoak izango liratekeela argi edukita. Horretaz gain, sarreran aipatu bezala, kontuan hartu behar da IAn gaur egun aplikatzen den atzeranzko itzulpenari esker, bi hizkuntzen arteko itzulpena beti izango dela hobea baliabide gutxien dituen hizkuntzatik baliabide gehiago dituenetik baino; beraz, euskarazko sortzaileek aukera handiagoak izango dituzte beren edukiak baliabide handiagoko beste hizkuntza batzuetara modu egokian itzuliak izateko, alderantzizko zentzuan baino.

Osasun arloari dagokionez, tesi hau garatu bitartean Osakidetzako langile batzuek idatzitako hainbat iritzi-artikulu publikatu dira Itzulbide proiektuaren eta inplementatu beharreko itzultzaile automatikoaren kontra,[6, 7] proiektua gazteleradun elebakarren mesedetan aurrera atera dela eta bere helburua txosten klinikoak euskaraz idaztea galaraztea dela iritzita. Tesi-egilearen iritzitan, kezka hauek ez dute oinarri sendorik, hasieratik euskaratik gaztelerara itzultzea lehenetsi delarik eta corpus bilketari esker betidanik gazteleraz idatzi duten osasun-langileek txosten klinikoak euskaraz idazteko pausoa eman dutelarik. Txosten klinikoak euskara hutsean idazteak suposatzen duen desobedientzia ariketa ontzat hartuta ere, jendarte elebidun batean kokatzen den osasun-sistema publiko batean euskaldunen eskubide linguistikoak babesteaz gain, pazientearen segurtasuna bermatzea beharrezkoa da, eta tesi honetan aurkeztutako irtenbideak (IA + post-edizioa) bi baldintza horiek betetzen ditu.

Edonola ere, itzultzailea Osakidetzan euskararen erabilera sustatzen laguntzeko asmoa du eta inola ere ez ditu ordeztu nahi bestelako euskararen alde egin beharreko ekimenak.

Azkenik, tesiaren edukian murgildu baino lehen, beronen mugak modu argi eta zehatzean aipatu nahi dira:

1. Garatutako itzultzaile automatikoa Osakidetzan inplementatzea helburu izanik, euskararen eta gazteleraren arteko itzulpena du helburu, gure artean erabiltzen den frantsesa albo batera utziz. Azken hilabeteetan frantziar estatuko mediku euskaldunak beren lana euskaraz egin ahal izateko antolatzen hasi direlarik,[8] etorkizuneko lan bezala uzten da eus-

[6] https://www.berria.eus/paperekoa/1896/022/001/2019-09-14/idiak-atzetik.htm

[7] https://www.berria.eus/paperekoa/2084/025/002/2021-01-08/euskara-ez-da-oztopo-bat-izan-behar.htm

[8] https://iparraldekohitza.eus/2021/06/04/maia-lacroix-ama-hizkuntzan-

kararen eta frantsesaren arteko osasun arloko itzultzaile automatiko bat garatzeko aukera, betiere bertako profesionalekin elkarlanean.

2. Garatutako itzultzaile automatikoak testua idatzizko forman jaso eta sortzen du, beraz ez da gai hizketa edo zeinu-hizkuntza interpretatu eta itzultzeko. Kontuan hartuta ohiko eszenatoki bat dela haur euskaldun elebakarrak mediku gazteleradun elebakarrekin komunikatu ahal izateko gurasoek itzultzaile lanak egin behar izatea, etorkizuneko lan bezala lehentasunez kontsideratuko da garatutako itzultzaile automatikoa ahozko formara hedatzea, beharrezkoak diren pribatutasun neurriak hartuz.

3. Erabilitako txosten klinikoen pribatutasuna mantentzeko neurrietako bat jasotako dokumentuetan esaldiak ausaz berrordenatzea izan da, garatutako itzultzaileak dokumentu mailan entrenatzea ezinezko eginez. Beraz, itzultzailea esaldi mailan entrenatu eta ebaluatu da, esaldi mailaz haratagoko fenomeno linguistikoen ebaluazioa oztopatuz.

4. Azkenik, garatutako itzultzaile automatikoa genero estereotipoak erreproduzitzen dituen corpus batean entrenatua izanik, honen eragina murrizteko neurriak adostuko dira Osakidetzarekin elkarlanean, izan corpusa bera moldatzeko zein garatutako sistema zuzentzeko.

1.6 Tesi-txostenaren egitura

Tesi hau zazpi kapituluz osatua dago. Sarrera den 1. kapitulu honen ostean, tesi honen oinarri diren aurrekariak azalduko ditugu (2. kapitulua); bai itzulpen automatikoari dagokionean, baita terminologia klinikoaren euskaratze automatikoari dagokionean ere. Bestalde, garatutako sistema batzuetan aplikatu den datuen hautespena (jatorrizko ingelesez, *data selection*) teknika azalduko da, eta atzeranzko itzulpenen aniztasun lexikala neurtzeko erabilitako metrikak deskribatuko dira.

Hurrengo kapituluan Itzulbide proiektua aurkeztuko dugu (3. kapitulua). Hasteko, tesi honen hasieran geneuzkan erronkak zerrendatuko ditugu, ondoren proiektuan bertan finkatutako helburuekin lotzeko asmoz. Atal honetan deskribatuko dira jasotako txosten kliniko elebidunen sailkapena, eta proiektuaren barruan corpusa bera biltzeko garatutako web-aplikazioa.

`aritzean-osasun-arreta-hobetzen-da/`

Jarraian, itzultzaile automatikoa entrenatu eta ebaluatzeko erabilitako baliabide guztiak deskribatuko dira 4. kapituluan. Alde batetik, erabilitako corpusak deskribatuko ditugu, hauek 4 taldetan banatuz: 1) domeinuz kanpoko corpus elebidunak, 2) terminologia kliniko eleanitza, 3) domeinu klinikoko corpus elebidunak, eta 4) domeinu klinikoko gaztelerazko corpusak. Bestalde, erabilitako itzultzaile automatiko desberdinak zerrendatuko dira, erabiltzen duten teknikaren arabera (EOIA, IAE edo IAN) sailkatuta, eta atzeranzko itzulpena egiteko edo azken helburu den euskara - gaztelera (aurrerantzean, eu-es) norabiderako erabili diren aipatuz. Kapitulu honetan deskribatuko da ere corpusei aplikatutako aurreprozesua.

Ostean, 5. eta 6. kapituluetan tesi honen muina diren metodologia eta emaitzak aurkeztuko dira, lehenengoan domeinuko corpusik gabeko egoeran aurrera eramandako lanak aipatuz eta bigarrenean domeinuko corpus elebidunekin egindako esperimentuak azalduz. Kapitulu hauen barne-egiturak egindako esperimentuen orden kronologikoa jarraitzen du, atal bakoitzean aztertutako ikergaiak deskribatuz.

Ondoren, 7. kapituluan bildutako emaitzen ondorioak eta ekarpenak zerrendatuko dira, aurretik esperimentu bakoitzaren ondorioetan zenbakituak izan direnak. Kapitulu berean, itzultzaile automatikoa Osakidetzan inplementatu ondoren egiteke geratu diren etorkizuneko lanak aipatuko ditugu. Hemen azalduko dira, besteak beste, genero alborapena murrizteko planteatzen diren irtenbideak, zein itzultzaile automatikoa beste eremu batzuetara (hizkuntza pareak, hizketa, etab.) zabaltzeko egon daitezkeen aukerak.

1.7 Argitalpenak

Tesi honetan zehar hainbat lan argitaratu ditugu, bai aldizkaritan, bai kongresuetan, baita liburuetan ere. Argitalpen guztiak ingelesez idatziak izan dira.

Aldizkarietan argitaratutako artikuluak

- Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. **Neural machine translation of clinical texts between long distance languages**. *Journal of the American Medical Informatics Association* 26(12), 1478–1487. orr. 2019.

Kongresuetako argitalpenak

- Soto X., Perez-De-Viñaspre O., Oronoz M., Labaka G. **Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish**. *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation (at MT Summit 2019)*, 8–18. orr. Dublin, Irlanda. 2019.

- Soto X., Shterionov D., Poncelas A., Way A. **Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation**. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, 3898–3908. orr. (*online*). 2020.

- Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. **Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation**. *Proceedings of the 5th Conference on Machine Translation (WMT2020)*, 875–880. orr. (*online*). 2020.

Liburuetan argitaratutako artikuluak

- Soto X., Perez-De-Viñaspre O., Oronoz M., Labaka G. **Development of a Machine Translation system for promoting the use of a low resource language in the clinical domain: the case of Basque**. *Natural Language Processing in Healthcare: A Special Focus on Low Resource Language*. Argitaratzeko zain.

Errebisio prozesuan

- Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. **Comparing and combining tagging with different decoding algorithms for back-translation in NMT: an analysis from a lexical diversity perspective**. Errebisio prozesuan.

2. KAPITULUA

Aurrekariak

Kapitulu honetan tesian zehar landuko ditugun arlo desberdinen oinarriak deskribatuko dira. Tesia itzulpen automatikoaren inguruan izanik, kapitulu honetako eduki gehiena horri buruz izango da. Horrela, lehen azpiatalean (2.1. atala) IA teknika desberdinak deskribatuko dira, sare neuronalen arkitektura desberdinak aipatuko dira eta erabiliko diren ebaluazio-metrika desberdinak zerrendatuko dira. Ondoren, garatutako itzultzaile automatikoan erabilitako SNOMED CTren euskaratzea nola egin zen azalduko dugu (2.2. atala), azken bi atalak tesi honetan zeharka landutako datuen hautespena (2.3. atala) eta aniztasun lexikala (2.4. atala) deskribatzeko erabiliko direlarik.

2.1 Itzulpen Automatikoa

Itzulpen automatikoa (IA) testu bat hizkuntza batean emanda, beste hizkuntza batean esanahi bera duen testu bat sortzea helburu duen ataza da.

IA egiteko lehenengo proposamen zehatzak 1933. urtean iritsi ziren, George Artsrouni eta Petr Smirnov-Troyanskik modu independentean erregistratutako patenteen bidez (Hutchins, 1995). Ordea, IAren ideia ez zen publiko orokorrera iritsi 1949. urtean Warren Weaverrek bere *memorandum*-a idatzi zuen arte (Weaver, 1949). Honen ondoren, Amerikar Estatu Batuetako hainbat unibertsitateetan IA ikertzen hasi ziren.

Ondorengo urteetan, Amerikar Estatu Batuak eta Sobietar Batasuna-

ren arteko gerra hotzaren testuinguruan, ingelesaren eta errusieraren arteko itzulpena garatzen hasi zen, gehienbat testu zientifiko eta teknologikoetan zentratuz. Garai horretan IA ataza ebaztea erraza izango zela uste zen, baina urte batzuk geroago ikusi zen ez zela berehalakoa izango. Horrela, 1966. urtean argitaratutako ALPAC txostenean (ALPAC, 1966), inbertitutako diru kopuru handiak nahikoa etekinik eman ez zuela ondorioztatu zen, etorkizunean inbertsioak murriztea proposatuz. Honek Amerikar Estatu Batuetan IAren inguruan egindako garapena nabarmen moteldu zuen.

Ordutik, hainbat teknika desberdin garatu dira IA egiteko, garai bakoitzean gailendu direnak. Hasiera batean, bi hizkuntzen artean testuak itzultzeko planteatu zen modua hiztegien bidez hitzez hitz itzultzeko erregelak sortu eta jatorri- eta helburu- hizkuntzetako gramatikak kontuan hartuz sortu beharreko testua moldatzea izan zen.

Ordea, 1990eko hamarkadan, digitalizatutako testu elebidunen kopurua handitzen joan zen heinean, corpusetan oinarritutako teknikak gailentzen joan ziren, hasierako Itzulpen Automatiko Estatistikotik hasi, eta azken urteetan baliabide konputazionalen handitzeari esker nagusitu den Itzulpen Automatiko Neuronaleraino.

Itzultzeko erabilitako teknika edozein izanda ere, bi ezaugarri dira itzultzaile automatiko bat ebaluatzerakoan kontuan hartzen direnak: 1) zehaztasuna, jatorri- eta helburu- hizkuntzetako esaldien esanahien arteko antzekotasuna neurtzen duena; eta 2) naturaltasuna, sortutako testuaren jarraikortasuna edo ulerkortasun maila neurtzen duena.

Atal honen amaieran, itzultzaile automatikoak ebaluatzeko erabili ohi diren metrika automatikoak definituko ditugu, giza ebaluazioetan aurrera eramaten diren prozedurak definitzearekin batera.

Jarraian, IAko 3 teknika nagusiak (EOIA, IAE eta IAN) labur deskribatuko ditugu, bakoitzaren abantaila eta desabantailak aipatuz.

2.1.1 Erregeletan Oinarritutako Itzulpen Automatikoa

Sistema hauek hizkuntzalariek definitutako erregeletan oinarritzen dira, eta esaldi bakoitza modu determinista batean itzultzen dute, itzulpenaren domeinua edozein delarik ere. Horretarako, jatorri-hizkuntzako termino bakoitza helburu-hizkuntzako terminoekin parekatzen duen lexikoi elebidun bat erabiltzen dute, hizkuntza bakoitzeko testua interpretatu zein sortzeko beharrezkoak diren erregela gramatikalak barneratuz.

Hau egiteko, beren funtzionamendua 3 fasetan banatzen dute: analisia,

transferentzia eta sorkuntza. Lehenengo fasean jatorri-hizkuntzako testua aztertzen da, hitz bakoitzaren lema eta informazio morfologikoa erauziz, *chunk* edo hitz-segidak identifikatuz, eta hauen arteko eta barneko erlazio sintaktikoak zehaztuz. Bigarrenean, informazio hau guztia helburu-hizkuntzara transferitzen da, lexikoi elebidunen bidez ordaina duten hitzak itzuliz eta bestelako informazio estrukturala erregelen bidez helburu-hizkuntzara moldatuz. Azkenik, hirugarren fasean helburu-hizkuntzako testua sortzen da, jatorri-hizkuntzatik transferitutako informazio sintaktiko eta morfologikoa erabiliz, horretarako espresuki definitutako erregelen bidez.

EOIAren abantaila nagusiak entrenatzeko corpusik behar ez izatea eta sarrerako testua edozein izanda ere irteerako testuren bat sortzea dira. Trukean, hizkuntzak berezkoa duen anbiguotasuna lantzeko zailtasuna eta garapen zein mantenurako lan handia behar izatea dira sistema hauen desabantaila nagusiak.

2.1.2 Itzulpen Automatiko Estatistikoa

Itzultzaile estatistikoek corpus elebidun zein elebakarretatik modu automatikoan ikasten dituzte aplikatu beharreko itzulpen-erregelak. Hasiera batean sistema estatistikoek unitate bezala hitzak erabiltzen bazituzten ere, aurrerago ikusi zen testuingurua kontuan hartzeko beharrezkoa zela hitzen ordez hitz-segidak aztertzea. Hortaz, itzultzaile estatistikoen oinarria corpusetan hitz-segida bakoitzaren agerpen probabilitateak neurtu, eta horren arabera jatorrizko esaldi berri bati dagokion itzulpen probableena bilatzean datza.

Prozesu horretan 3 modulu desberdin sortzen dituzte: corpus elebidunetik ikasitako itzulpen-eredua eta berrordenatze-ereduak, eta helburu-hizkuntzako corpus elebakarretik ikasitako hizkuntza-eredua. Itzulpen-ereduak jatorri- eta helburu-hizkuntzen arteko hitz edo hitz-segida baliokideak identifikatzen ditu; berrordenatze-ereduak sortutako testua jatorri-hizkuntzatik helburu-hizkuntzara moldatzeko beharrezkoak diren berrordenatzeak ikasten ditu; eta hizkuntza-ereduak helburu hizkuntzako testua sortzerakoan hurrengo hitz probableena zein izango den adierazten du. Behin modulu hauek entrenatuta, sarrerako testu berriak itzuli ahal dira ikasitako erregelak aplikatuz.

Aurreko sistemekiko IAEren abantaila nagusia itzulpen-erregelarik definitu behar ez izatea da, eta emaitzei dagokienez hobekuntza handiena itzulpenen zehaztasunean eman zen. Ordea, esaldiak zatika itzultzeak sortutako testu batzuen naturaltasuna mugatua izatea ekarri zuen.

2.1.3 Itzulpen Automatiko Neuronala

Itzultzaile neuronalek ere corpusetatik ikasten dute nola itzuli, hasieran sare neuronal batek jatorrizko esaldiaren esanahia modu abstraktuan kodetuz, eta ondoren beste sare neuronal batek dekodetze-prozesuan helburu-hizkuntzako testua sortzen ikasiz. Hau da, kodetze-dekodetze hurbilpena erabiltzen dute. Horretarako, hitzak zein esaldiak zenbakizko bektore moduan errepresentatzen dira, bakoitzaren balioak modu automatikoan ikasten direlarik, ikasketa-prozesuan zehar sare neuronalen parametroen balioak moldatuz.

IAErekin alderatuta, IANak muturretik muturrera ikasten du, jatorri- eta helburu-hizkuntzetako esaldiak zuzenean erlazionatuz, modulu desberdinak entrenatu behar izan gabe. Ostera, IANak ere probabilitateekin funtzionatzen du, esaldi berri bat hitzez hitz sortzerakoan sarrerako esaldia eta ordura arte sortutako esaldiaren zatia kontuan hartuta probabilitate altuena izango lukeen hurrengo hitza sortuz.

IANaren abantaila nagusia corpus elebidunak emanda itzulpen-prozesua modu zuzenean ikastea da, IAEren zehaztasuna mantentzeaz gain, itzulpenen naturaltasunaren emaitzak nabarmen hobetuz. Desabantailarik handiena itzulpena nola egin duen ezin jakin edo interpretatzea da, behin garatuta topatutako akats espezifikoren bat zuzentzeko modulu zehatzen bat hobetzea ezinezko eginez.

IANaren barruan sare neuronalen arkitektura desberdinak definituak izan dira, bakoitzak bere abantailak eta desabantailak dituztelarik. Gaur egun, bi dira erabiltzen diren arkitektura nagusiak: Sare Neuronal Errekurrenteak (SNE) eta Transformer (Vaswani *et al.*, 2017). Lehenengo IAN sistemek SNEak erabiltzen zituzten (Kalchbrenner eta Blunsom, 2013; Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015), baina gaur egun Transformer arkitektura da arloan gailentzen dena.

SNEek jatorrizko esaldia hitzez hitz irakurtzen dute esaldiko ordenean, ikasitako esaldiaren errepresentazio bektoriala pauso bakoitzean moldatuz. Honen ondorioz, jatorrizko esaldiaren errepresentazioek esaldiko lehen hitzarekiko dependentzia handia dute, eta sistema hauek zailtasunak dituzte esaldi luzeak edo hitzen arteko distantzia handia duten erlazioak barnebiltzen dituzten esaldiak itzultzeko. Honi aurre egiteko, oinarrizko inplementazioan sare neuronaleko unitate bakoitzari memoria moduko bat gehitzea proposatu izan da, hasiera batean LSTM (jatorrizko ingelesez, *Long Short-Term Memory*) (Hochreiter eta Schmidhuber, 1997) izeneko neurona moten bitartez, eta ondoren, hauen sinplifikazio diren GRU (jatorrizko ingelesez, *Gated*

Recurrent Units) (Cho *et al.*, 2014) unitateen bidez.

Edonola ere, SNEek jatorrizko esaldiaren esanahia bektore bakarrean kodetu behar dute, esaldi horren hitz kopurua edozein dela ere. Muga hau gainditzeko, atentzio-mekanismoa sortu zen (Bahdanau *et al.*, 2015), zeinak sarrerako esaldiaren errepresentazio osoa erabili beharrean, esaldia osatzen duten token bakoitzaren errepresentazioa gordetzea ahalbidetzen duen. Ondoren, dekodetze garaian, sistemak hitz berri bat sortu aurretik, hitz hori sortzeko esanguratsua den jatorrizko testuingurua kalkulatzen du, hitz horrekiko erlazio handiena duten sarrerako esaldiko hitzak aukeratuz eta beraien banakako errepresentazioak konbinatuz.

Transformer arkitekturan, ordea, esaldi bat kodetu edo dekodetzean hitz bakoitzak beste hitz guztiekiko dituen erlazio edo dependentziak ikasten dira, aurretik aipatutako atentzio-mekanismoan oinarrituz. Honek sistemei aukera ematen die erlazio linguistiko konplexuak ikasi eta modu egokian itzultzeko, erlazio hori osatzen duten hitzen arteko distantzia edozein izanda ere.

Arkitekturaz gain, hainbat dira sare neuronalak definitzen dituzten ezaugarri edo hiperparametroak. Horien artean, sarearen geruza kopurua, geruza bakoitzeko neurona kopurua, hitz bakoitzaren esanahia kodetzeko erabili den *embedding* edo hitz-bektorearen tamaina; ikasketa-prozesuan erabilitako optimizatzailea eta *batch* edo esaldi multzoaren tamaina; eta dekodetze-prozesuan momentu bakoitzean gordetako irteera posibleen kopurua (jatorrizko ingelesez, *beam width*). Ikusiko dugun moduan, hiperparametro hauen balioek ere eragina izan dezakete lortutako itzulpenen kalitatean (Britz *et al.*, 2017). Tesi-lan honetan, adibidez, optimizatzaile moduan Adadelta (Zeiler, 2012) eta Adam (Kingma eta Ba, 2014) erabili dira.

2.1.4 IAren artearen egoera

Gaur egun, IAren artearen egoera definitzeko eremu nagusia WMT konferentzia da, eta bertan jorratzen diren ataza desberdinetatik albisteen itzulpena da parte-hartzaile gehien izan ohi dituena. Bestalde, lantzen diren hizkuntza pare desberdinen artean, ingelesa - alemana eta ingelesa - frantsesa dira IAren artearen egoera neurtzeko erreferentzia bezala erabili ohi diren hizkuntza pareak.[1] Horietan lortzen diren emaitzen arabera, IAN da gaur egun IA egiteko gailentzen den teknika, eta Transformer da emaitza hoberenak

[1] `https://github.com/sebastianruder/NLP-progress/blob/master/english/machine_translation.md`

lortzen dituen arkitektura. Orokorrean, IAN sistemak zenbat eta parametro gehiago eduki (geruza edo neurona kopuru handiagoaren bidez), emaitzak are eta hobeagoak izango dira. Ondorio hauek hizkuntza pare eta domeinuaren arabera alda daitezkeen arren, modu orokorrean interpreta daitezke.

Hainbat ekarpen izan dira IAN gainontzeko teknikei nagusitzea ahalbidetu dutenak, horietatik hitzen segmentazioa (Sennrich *et al.*, 2015) eta atzeranzko itzulpena (Sennrich *et al.*, 2016) nabarmentzen direlarik.

Hitzen segmentazioaren bitartez, corpusetan oinarritutako tekniken muga bati aurre egiten zaio. Izan ere, sistema hauek mugatuak daude interpreta ditzaketen hitz kopuruari dagokionean, hiztegiaren tamaina handitzeak dekodetze-abiadura nabarmen moteltzen duelarik. Arazo honi aurre egiteko, informatikan konpresio-metodo bezala ezaguna zen BPE teknika IAra egokitu zen (Sennrich *et al.*, 2015), prozesuaren abiapuntuan hitzak karakteretan banatuz eta ikasketa-prozesuko pauso bakoitzean corpuseko karaktere edo karaktere multzo ohikoenak karaktere multzo berri batean batuz. Teknika hau bereziki lagungarria da euskararen antzera morfologikoki aberatsak diren hizkuntzetarako. Hitzen segmentazioa IAn aurreprozesuaren azken pausoa izan ohi da, eta bere aldagai nagusia definitzen den iterazio kopurua edo hiztegiaren tamaina da. Aurrerago, jatorrizko BPEren gainean erregularizazioa aplikatzea proposatu da (Provilkov *et al.*, 2020), BPEren baliokidea den *sentence-piece*[2] metodoan erregularizazio hau berezkoa delarik.

Atzeranzko itzulpenari dagokionez, ideia sinple batean oinarritzen da: definitutako hizkuntza parerako, garatu nahi den itzulpen noranzkoaren kontrako zentzuan sistema bat diseinatu, eta horrekin helburu-hizkuntzako corpusak jatorri-hizkuntzara itzuli, azken sistema entrenatzeko eskuragarri dagoen corpusaren tamaina modu artifizialean handituz. Honek garatutako sistemen kalitatea nabarmen hobetzea ahalbidetzen du, helburu-hizkuntzan egon daitezkeen corpusei etekina ateraz.

Modu orokor batean aplikatzen diren teknika hauetaz gain, badira eszenatoki espezifikoetara moldatzen diren metodoak. Horietatik azpimarragarriena, testu elebidunik erabili gabe itzulpena gauzatzeko gai den itzulpen automatiko gainbegiratu gabea da (Artetxe *et al.*, 2017; Lample *et al.*, 2017). Hurbilpen honetan, jatorri- eta helburu-hizkuntzetako corpus elebakar bakoitzeko hitz desberdinen hitz-bektoreak modu independentean ikasten dira, ondoren mapaketa baten bitartez bi hizkuntzen espazio bektorialak parekatuz, eta hitz bakoitzaren itzulpena egiteko hitz horri dagokion hitz-bektoretik

[2] `https://github.com/google/sentencepiece`

beste hizkuntzan gertuen dagoen hitz-bektorea aukeratuz.

Praktikan, ordea, eszenatoki ohikoagoa da ataza definitzen duen hizkuntza pare edo domeinurako esaldi pare elebidun gutxi batzuk izatea, eta esaldi pare elebidun gehiago dituen beste hizkuntza pare edo domeinu bateko corpusa oinarri bezala erabiltzea. Hurbilpen hau *transfer learning* edo transferentzia bidezko ikasketa terminoarekin ezagutzen da (Zoph *et al.*, 2016), eta bere adierazpiderik sinpleenean sistema bat baliabide handiko corpus batean entrenatzen da, ondoren azken helburu den hizkuntza pare edo domeinuko esaldiekin entrenamendua jarraituz.

Orokorrean, itzulpen-sistema bat domeinu batera egokitu nahi dugunean bi hurbilpen mota daude: 1) datuetan oinarritutako egokitzapena, eta 2) ereduetan oinarritutako egokitzapena (Chu eta Wang, 2018). Lehenengoan, sistemaren arkitektura aldatu gabe datu berriak txertatzen dira itzultzailea domeinu horretara egokitzeko; bigarrenean, ereduaren arkitektura bera moldatzen da datu gutxi dituzten domeinuetan itzulpen automatikoaren kalitatea hobetzeko.

Domeinura egokitzeko proposaturiko metodoen artean, Hu *et al.* (2019) lanean IAN sistemak birdoitzeko atzeranzko itzulpena hitzez hitz aplikatzea proposatzen dute, horretarako domeinuko corpus ez paraleloan gainbegiratu gabeko lexikoaren indukzioa eginez. Halaber, Chu *et al.* (2017) artikuluan hurbilpen sinpleago bat proposatzen dute, birdoitzerakoan domeinuko corpusa bakarrik erabili ordez, hau domeinuz kanpoko corpusarekin nahastuz, tamaina desberdintasuna orekatzeko domeinuko corpusa hainbat aldiz errepikatuz. Horretaz gain, emaitzak hobetzeko etiketak erabiltzen dituzte domeinuko eta domeinuz kanpoko corpusak desberdintzeko. Khayrallah *et al.* (2018) lanean ordea, erregularizazioa erabiltzen dute IAN sistemak domeinura egokitzeko, birdoitze-prozesuan sistemaren entrenamendu-helburuan termino berri bat gehituz, domeinuko ereduaren eta domeinuz kanpoko ereduaren hitzen distribuzioen arteko entropia gurutzatua minimizatzeko.

Bestalde, itzultzaile automatiko bat domeinu batera egokitzeko edo bezeroaren beharretara egokitzeko ohikoa izaten da terminologia modu zehatz batean itzuli nahi izatea. Horretarako, bi metodo nagusitzen dira: 1) dekodetze-prozesuan sortzen diren hitzak aurretik definitutako hiztegi baten arabera mugatzea, dekodetzea bera motelduz (Hokamp eta Liu, 2017); eta 2) entrenamendu-corpusa moldatzea, sistemak bere kabuz ikas dezan markatutako terminoak modu zehatz batean itzuli behar direla (Dinu *et al.*, 2019). Metodo hauek IAN sistemen itzulpena nolabait kontrolatzeko aukera ematen dute, baina trukean testuinguruaren arabera desegokiak izan daitezkeen

itzulpenak sor ditzakete.

Domeinu klinikoko IAren garapenari dagokionez, erreferentziak eskasak dira, ziurrenik domeinuko datuen pribatutasunarengatik eta arloak eskatzen duen kalitate altuarengatik. Aurki daitezkeen lehen erreferentzien artean, adibide moduan Liu eta Cai (2015) sistema estatistikoa aipatu dezakegu, non txosten klinikoak ingelesaren eta gazteleraren artean itzultzeko garatutako sistemaren kalitatea sarean eskuragarri dauden sistemenarekin alderatzen duten. Sistema zehatzetaz haratago, nabarmentzekoak dira osasun arloko IAn egindako lanak biltzen dituzten berrikuspen-artikuluak. Horien artean lehenengoan, Dew *et al.* (2018) artikuluan ordura arte garatutako sistemak aztertzen dituzte, gehienak EOIA eta IAE sistemak izanik. Sistemen kalitatea eta osasun arloak beharrezkoa duen zehaztasuna kontuan hartuta, sortutako itzulpenak post-editatu beharra azpimarratzen dute. Beranduago, Haddow *et al.* (2021) itzultzaile profesionalei zuzendutako liburuan osasun arloko IAri buruzko kapitulua osatzen dute, orain arte egindako aurrerapenak barnean hartuz. Bertan, besteak beste, baliabide gutxiko hizkuntzetan egindako MeMat proiektu pilotua deskribatzen da, zeinaren helburua Hegoafrikako ospitaleetan medikuak isiXhosa hizkuntza bakarrik ezagutzen duten pazienteekin komunikatzeko itzultzaile automatiko bat garatzea izan zen. Corpusa txikia izanda lortutako emaitzak baxuak izan baziren ere, baliabide gutxiko hizkuntza eta domeinuetarako transferentzia bidezko ikasketa lagungarria izan zitekeela ondorioztatu zuten.

Azkenik, nahiz eta lantzen den arloa desberdina eta zabalagoa izan, aipatzekoa da WMT konferentzian biomedikuntzaren arloko testuak itzultzeko antolatzen den ataza partekatua (Bawden *et al.*, 2020). Bertan, artikulu zientifikoen laburpenetatik erauzitako esaldiak itzultzeko sistemak garatzen dira, eta azken bi urteetan ingelesa - euskara hizkuntza parea ere gehitu da.

2.1.5 IAren ebaluazioa

HPko ataza guztietan bezala, IAren ebaluazioa egiteko eszenatoki optimoa gizakiek egindako ebaluazioa litzateke, ahal izanez gero domeinuaren ezagutza duten itzultzaile profesionalek egindakoa izanik. Honek, ordea, itzultzaile automatikoen garapena moteldu eta garestitzen du, praktikan ezinezkoa baita sisteman aldaketa bat egiten den bakoitzean giza ebaluazio bat egitea.

IAn aurrerapenak azkartzeko, itzultzaileen kalitatea estimatzeko balio duten metrika automatikoak erabiltzen dira. Hauek sistemen irteera eta erreferentziazko itzulpenak emanda modu azkar batean sistemen kalitatearen

neurgailu diren zenbaki batzuk ematen dituzte. Definitutako metrika gehienek muga nabarmen bat dute: esaldi bat itzultzeko modua ez dela bakarra. Honi aurre egiteko badira erreferentziazko itzulpen anitzak kontuan hartzeko gai diren metrikak, baina honek ere ebaluazioaren kostua handitzen du, normalean erreferentziazko itzulpenak itzultzaile profesionalek eginak direlako.

Gaur egun, BLEU (Papineni *et al.*, 2002) da gehien erabiltzen den metrika (Marie *et al.*, 2021), nahiz eta giza ebaluazioarekin duen korrelazioa baxua dela frogatua izan (Callison-Burch *et al.*, 2006). Azkenaldian, ordea, komunitate zientifikoan aurre-entrenatutako metrikak (adib.: COMET Rei *et al.*, 2020) erabiltzea gomendatzen hasi da (Kocmi *et al.*, 2021), itzulpen-sistemen kalitatearekiko adierazkorragoak direlakoan. Aldiz, aurre-entrenatutako metriken desabantaila nagusia hizkuntza pare guztietarako baliagarriak ez izatea da, esanguratsuak izan ahal izateko corpus erraldoietan entrenatu behar direlako, zeinak gaur gaurkoz munduko hizkuntza nagusietarako bakarrik eskuragarri dauden.

Orokorrean, ohiko metriken mugei aurre egiteko sistema desberdinak konparatzeko ebaluazio-metrika bat baino gehiago erabiltzea gomendatzen da. Jarraian, tesi honetan erabili diren metrikak deskribatzen dira. Erreferentzia bezala, Haddow *et al.* (2021) osasun arloko itzulpen automatikoari buruzko liburu-kapituluan ebaluazio-metrika hauek aipatzen dira ohikoenak bezala.

- **BLEU** (jatorrizko ingelesez, *BiLingual Evaluation Understudy*) (Papineni *et al.*, 2002): funtsean, zehaztasunean oinarritutako metrika bat da, erreferentziazko itzulpenean eta itzultzailearen irteeran aldi berean agertzen diren hitz-segidak zenbatuz kalkulatzen dena. Normalean, 4 hitzetarainoko hitz-segidak kontuan hartzen dira, eta irteera laburregiak penalizatzen dira sistemak ohikoen diren hitzak bakarrik sortzea zigortzeko. Bere balioa 0 eta 100 artean dago, eta balioa zenbat eta handiagoa izan itzultzailearen kalitatea handiagoa izango dela estimatzen da.

- **TER** (jatorrizko ingelesez, *Translation Edit Rate*) (Snover *et al.*, 2006): metrika honek sistemaren irteeratik erreferentziazko itzulpenera iristeko egin beharreko aldaketa kopuruak zenbatzen ditu, bere balioak normalizatzeko erreferentziazko itzulpenen luzeren bataz bestekoa erabiliz. Honek ere 0 eta 100 arteko balioak hartzen ditu, baina kasu honetan balioa zenbat eta txikiagoa izan, hainbat eta handiagoa izango da sistemaren kalitatearen estimazioa.

- **METEOR** (jatorrizko ingelesez, *Metric for Evaluation of Translation with Explicit ORdering*) (Banerjee eta Lavie, 2005): hasiera batean, erreferentziazko itzulpena eta sistemaren irteeran, bietan, agertzen diren hitzak zenbatzen ditu. Ondoren, parekatu gabeko hitzak lematizatu eta berriro konparatzen ditu, parekatze berriak ere zenbatuz. Azkenik, penalizazioak ezartzen ditu erreferentziazko itzulpena eta itzultzailearen irteeraren artean hitzen ordena aldatu behar izan bada. BLEUren modura, 0 eta 100 arteko balioak har ditzake, eta zenbat eta handiagoa izan METEOR balioa itzultzailearen kalitatea ere handiagoa izango dela estimatzen da.

- **chrF** (jatorrizko ingelesez, *character n-gram F-score*) (Popović, 2015): metrika honek itzulpenen kalitatea estimatzeko karaktereak erabiltzen ditu hitzen ordez, eta zehaztasuna neurtzeaz gain, estaldura ere kontuan hartzen du, karaktere-segida guztien bataz besteko *F-score*-a edo F-neurria kalkulatzen duelarik. Defektuz, 6 karaktereetarainoko luzera duten karaktere-segidak kontuan hartzen ditu, eta guk erabilitako aldaeran F3-neurria kalkulatu dugu. Bere balioen interpretazioa BLEU eta METEOR-ena bezalakoa da, eta azken ikerketen arabera metrika hau erabiltzea gomendatzen da aurre-entrenatutako metriken lagungarri edo ordezko bezala (Kocmi *et al.*, 2021).

Giza ebaluazioari dagokionean, ohikoena sortutako itzulpenen zehaztasuna eta naturaltasuna neurtzea da, horretarako ebaluatzaile bakoitzari esaldi eta aldagai bakoitza 1etik 5era doan eskala batean kokatzeko eskatuz (jatorrizko ingelesez *Likert scale* terminoarekin ezagutzen dena).

Zehaztasunaren barruan, errorea neurtzeko hainbat neurri kontsidera daitezke, horietatik ohikoenak gaizki itzulitako hitzen kopurua, jatorrizko esaldian agertu baina sistemak itzuli gabeko hitzen kopurua, eta sarreran agertu gabe itzultzailearen irteeran agertzen diren hitzen kopuruak direlarik.

Beste aukera bat ebaluatzaileei sortutako itzulpenak zuzendu edo post-editatzeko eskatzea da, horretarako behar duten denbora eta egindako aldaketen kopuruak sistemaren kalitatearen neurgailu bezala erabiliz. Honek denbora handiagoa eskatzen du, baina sistemak modu zehatzagoan ebaluatu ahal izateko aukera ematen du, sor daitezkeen akats motak identifikatuz.

Ebaluazioa egiterakoan bi aukera nagusi daude sistemaren irteera eta erreferentziazko testuak alderatzeko, baldin eta konparaziorako sistemari ematen zaion sarrera edo erreferentziazko itzulpena erabiltzen den. Orokorrean,

ebaluatzaileak jatorri- eta helburu-hizkuntzak ezagutzen baditu sistemaren sarrera eta irteerak konparatzea lehenesten da; alde batetik sistemak ezin baitu itzuli sarreran agertzen ez den hitzik, eta bestetik, aipatu dugun moduan baliozko itzulpen posibleak asko izan daitezkeelako.

Era berean, giza ebaluazioa izatez subjektiboa denez, esaldi bakoitza ebaluatzaile batek baino gehiagok ikuskatzea gomendatzen da, beraien arteko adostasuna ebaluazioaren kalitatearen neurgailu bezala adieraziz.

2.2 Osasun-alorreko terminoen euskaratzea

Sarreran aipatu moduan, txosten klinikoak euskaraz idatzi ahal izateko mugetako bat euskarazko terminologia klinikoa estandarizatu gabe egotea da. Aldi berean, txosten klinikoak euskaratik gaztelerara modu egokian itzuliko dituen sistema bat garatu nahi bada, beharrezkoa izango da sistema hori terminologia klinikoa ezagutzeko gai izatea.

Helburu bikoitz horrekin, tesi hau hasi baino lehen SNOMED CT datubase terminologikoa modu automatikoan euskaratzeko proiektua garatu zen (Perez-de Viñaspre, 2017), zeina tesi honetan diseinatutako itzultzailearen terminologia klinikoaren estaldura handitzeko baliagarria izango den.

SNOMED CT mundu mailako terminologia kliniko eleanitzik zabalena kontsideratzen da, eta jatorriz ingelesez idatzia izanik, euskaratzea aurrera eramaterakoan ingelesezko edukia hartu zen eredu moduan.

Euskarazko termino klinikoen sorkuntza lau fasetan definitu zen:

1. Lehenengo fasean hiztegiak erabiltzen ziren ingelesezko terminoei dagozkien euskarazko ordainak esleitzeko. Fase honetan Euskalterm, Zientzia eta Teknologiaren Hiztegi Entziklopedikoa, Giza Anatomiako Atlasa eta Erizaintzako hiztegiak erabili ziren, besteak beste.

2. Bigarren urratsean, domeinu klinikoan ohikoak diren latinetik eta grezieratik eratorritako termino neoklasikoak identifikatzen dira jatorrizko ingelesean, eta euskarazko termino baliokideak sortzeko transliterazioerregelak aplikatzen dira, garatutako *Neoterm* tresnaren bitartez.

3. Hirugarren pausoan, egitura zehatza duten hitz anitzetako termino klinikoak euskaratzeko gai den *KabiTerm* tresna garatu zen. Honi esker, termino habiaratuak (adib: *'fracture of elbow'*) '[GAIXOTASUN]

of [EGITURA]' txantiloiarekin identifikatu eta '[EGITURA]-ren [GAIXOTASUN]' egiturarekin euskaratzen dira (aurreko adibideari jarraituz, 'ukondoaren haustura' terminoa sortuz).

4. Azkenik, laugarren fasean aurreko faseetan itzuli ezin izan diren terminoak itzultzeko EOIA erabiltzen da, kasu honetan *Matxin* (Mayor, 2007) medikuntza-domeinura egokitzeko bere lexikoietan hiztegi klinikoak txertatuz. Horrela, *MatxinMed* tresna sortu zen.

Tesi honetan garatutako itzultzaile automatikoan erabili den terminologia kliniko gehiena SNOMED CTren euskaratzetik datorren arren, eskuragarri dagoen bestelako terminologia ere txertatu dugu gure sistemetan. Horien artean GNS-10 nazioarteko gaixotasunen sailkapenaren euskaratzea da nabarmenena, berau eskuz itzulia izan delarik.[3]

Azkenik, COVID-19aren inguruan sortutako terminologia berria ere txertatu nahian, tesian zehar euskaratu diren termino gutxi batzuk ere bildu ditugu. COVID-19aren inguruko terminologiaren jatorria 4.1. atalean deskribatuko dugu, terminologia guztien sarrera kopuruak ere atal horretan azaltzen direlarik.

2.3 Datuen hautespena

Orokorrean itzultzaile automatiko bat garatzerakoan erabiltzen den estrategia ahalik eta esaldi pare elebidun gehien lortzean zentratzen den arren, bada ikerketa-arlo bat, datuen hautespena izenekoa, corpus handietatik ebatzi nahi den atazara gehien hurbiltzen diren esaldiak aukeratzeaz arduratzen dena. Estrategia honi esker, sistemak entrenatzeko beharrezkoa den denbora, eta beraz kontsumitutako energia, murrizteko aukera sortzen da.

Datuen hautespena egiteko teknika desberdinak dauden arren, guk erabiliko duguna FDA (jatorrizko ingelesez, *Feature Decay Algorithms*) izenekoa da (Biçici eta Yuret, 2015; Poncelas *et al.*, 2018a), eta IArako egokiena dela frogatua izan da (Silva *et al.*, 2018).

FDA metodoa jatorrizko corpusetik garapen multzoko esaldiekiko antzekotasun handiena duten esaldiak aukeratzean datza. Horretarako, iterazio bakoitzean hautatu beharreko esaldietatik garapen multzoko esaldiekiko hitzsegida kopuru komunik altuena duen esaldia aukeratzen da. Era berean, hau-

[3] `https://drive.google.com/drive/folders/1gUQHoutvYIXGGPVTBbBF3qlHHhX9qbr0`

tatutako corpusa anitza izan dadin, esaldi berri bat aukeratzerakoan aurretik hautatutako hitz-segidak zigortu egiten dira.

Tesi honetan bi momentutan aplikatu dugu datuen hautespena: 1) IA teknika desberdinekin (EOIA, IAE eta IAN) atzeranzko itzulpena eginez sortutako corpus multzotik atazarako egokiena izan daitekeen azpimultzoa hautatzeko; eta 2) tesiko azken sistemak garatzerakoan, Osakidetzan urteetan zehar bildutako gaztelerazko corpus erraldoi bat atzeranzko itzulpenaren bidez euskarara itzuli eta sortutako corpus elebidunetik gure helburura hobekien egokitzen den esaldien azpimultzoa aukeratzeko.

Lehenengo esperimentuan helburu nagusia datuen hautespena atzeranzko itzulpenaren bidez sortutako corpusei aplikagarria zela frogatzea izan zen, bide batez datuen hautespen-prozesuak teknika desberdinen bidez itzulitako testuetan zer nolako eragina zuen aztertuz. Bigarrenean, aldiz, helburua praktikoagoa izan zen; Osakidetzako corpus elebakarra aurretik erabilitako corpus elebiduna baino handiagoa izanik, sistemaren kalitatea egokia izateko eta berau entrenatzeko denbora onargarria izateko beharrezkoa baitzen bere tamaina murriztea.

2.4 Aniztasun lexikala

IAN sistema batek esaldi berri bat sortzerakoan, sortu litezkeen esaldi posible guztietatik ikasitako probabilitate distribuzioan probabilitate altuena duen esaldia aukeratu behar du. Hau modu efizientean egiteko, jatorrizko ingelesez *beam search* (Tillmann eta Ney, 2003) izena duen dekodetze-teknika erabili ohi da, non momentu bakoitzean sortu litezkeen n esaldi probableenak bakarrik kontuan hartzen diren, n hori aurretik aipatutako *beam width* deritzona izanda.

Honek, eta ebaluazio-metrika gehienak zehaztasunean oinarrituak izateak, metrika horren balio altuena duten sistemak lehenestean sortutako itzulpenek jatorrizko testuek baino aniztasun lexikal txikiagoa izatea dakar, sistemak joera handiagoa izango duelako ohikoenak diren hitzak sortzeko (Vanmassenhove *et al.*, 2019).

Hau horrela izanda, atzeranzko itzulpena egiterakoan sortutako corpus elebidunaren muga nagusia azken sistemaren jatorri-hizkuntzako esaldiak lexikoki sinpleagoak izatea izango da, sistemaren estaldura murriztuz. Honek azaldu lezake atzeranzko itzulpena egiterakoan *beam search* ordez *sampling* edo laginketa erabiltzen denean (Edunov *et al.*, 2018) azken sistemaren emai-

tzak hobetzea.

Aniztasun lexikala neurtzeko metriketatik ezagunena TTR (jatorrizko ingelesez, *Type-Token Ratio*) (Templin, 1975) da, zeinak corpus bateko hitz mota desberdin guztien eta hitz kopuru totalaren arteko erlazioa kalkulatzen duen. Interpretatzeko erraza bada ere, bere balioak corpusaren tamainaren arabera asko alda daitezke, beraz, bere baliozkotasuna antzeko tamaina duten corpusen arteko konparazioetara mugatzen da.

Honi aurre egiten dion metriketako bat jatorrizko ingelesez *Yule's I* (Yule, 1944) deritzona da. *Yule's I* 'konstante karakteristiko' izenarekin ezagutzen den *Yule's K*-ren inbertsioa da, zeinak aztertutako testuaren maiztasun lexikalaren aldakortasuna neurtzen duen. Corpusen tamainan egon daitezkeen aldaketekiko egonkorragoa izan arren, bai TTR bai *Yule's I* corpus txikietara hobe egokitzen direla uste da.

Azkenik, MTLD (jatorrizko ingelesez, *Measure of Textual, Lexical Diversity*) (McCarthy, 2005) metrikak sekuentzialki neurtzen du TTR balio berdina duten hitz-segiden bataz besteko luzera. Modu sekuentzialean neurtua izanik, corpusaren tamainarekiko duen aldakortasuna txikiagoa da, eta IAn erabilitako corpus handietara hobeto egokitzen den metrika dela kontsideratzen da.

Tesi honetan atzeranzko itzulpenaren bidez sortutako corpusen aniztasun lexikala neurtu dugu, honek azken sistemen kalitatean izan dezakeen eragina aztertzeko. Horretarako, bi esperimentu multzo diseinatu ditugu: 1) teknika desberdinen bidez (EOIA, IAE eta IAN) sortutako corpusen aniztasun lexikala neurtzeko; eta 2) IANeko dekodetze-teknika desberdinak (*beam search*, *sampling*,...) erabilita sortzen diren corpusen aniztasun lexikala neurtzeko.

3. KAPITULUA

Itzulbide

Tesi hau hasterakoan, azken helburu den itzultzaile automatikoa garatzeko genuen oztopo nagusia domeinu klinikoko corpus elebidunen falta zen. Horregatik, Osakidetzari aurkeztutako Itzulbide proiektuan, itzultzaile automatikoa garatzeko konpromisoarekin batera, Osakidetzako langileek corpus elebidun bat biltzea aurreikusi zen.

Orokorrean domeinu batera egokitutako itzultzaile bat garatzeko beharrezkoa bada domeinu horretako corpus bat izatea, domeinu hori klinikoa bada, are beharrezkoagoa izango da. Izan ere, domeinu klinikoak hainbat ezaugarri ditu bestelako domeinuetatik aldentzen dituenak; eta gainera, maneiatzen den edukia pribatua izanda, zaila izango da sarean antzeko ezaugarriak dituen corpusik aurkitzea.

Edukiari berari dagokionean, terminologia aberatsa izatea da itzultzailearen garapenari zailtasunik handiena eransten diona, gure kasuan gainera euskarazko terminologia klinikoa estandarizatu gabe dagoela kontuan izanda. Horretaz gain, ezaguna da medikuek osasun-txostenetan erabiltzen duten hizkera ez estandarra, gramatika sinplifikatuak erabiliz, askotan aditzik edo puntuazio zeinurik gabeko esaldiak sortuz, laburdura espezifikoak erabiliz, eta azkar idaztearen ondorioz akats ortografikoak eginez.

Horregatik guztiagatik, txosten klinikoak modu egokian itzuli ahal izateko beharrezkoa da osasun-langileek idatzitako corpus errealista bat biltzea, goian aipatutako zailtasunak arintzeko aukerak emanez.

Kapitulu hau 4 ataletan banatzen da: 1.an Itzulbide proiektua bera deskribatuko da; 2.ean jasotako txostenak nola sailkatu diren adieraziko da;

3.ean corpusa biltzeko garatutako web-aplikazioa aurkeztuko da; eta 4.ean bildutako corpusaren xehetasunak emango dira.

3.1 Itzulbide proiektua

Itzulbide proiektua Osakidetzak aurkeztu eta Ixa taldeak irabazitako proiektu bat da, bere helburu nagusia txosten klinikoak euskararen eta gazteleraren artean itzultzeko gai den itzultzaile automatiko bat garatzea izanik.

Horretarako beharrezkoa da domeinu klinikoko corpus elebidun bat biltzea, eta helburu horrekin proiektuaren barruan txosten klinikoak euskaraz eta gazteleraz biltzeko web-aplikazioaren diseinua sartzen da.

Modu horretan, bai itzultzaile automatikoari esker, baita corpus elebiduna biltzeko saiakerari esker, txosten klinikoak euskaraz idatz daitezen sustatzen da, hau Osakidetzaren helburuetako bat izanik.

Izan ere, Osakidetzaren barnean geroz eta osasun-langile gehiago dira txosten klinikoak euskaraz idatzi nahiko lituzketenak, baina arestian aipatu bezala, egungo egoeran hau ezin da segurtasuna bermatuz praktikara eraman, langileen zati handi batek ez duelako euskara ulertzen.

Gainera, osasun-langileen arteko eta hauen eta pazienteen arteko ahozko elkarrizketak euskaraz izatea gero eta ohikoagoa bihurtzen ari da, modu naturalean osasun profesionalak euskaraz idaztera bultzatuz. Gaur egun, ordea, egoera hauek langileen partetik eskatzen den esfortzua handitzea dakarte, euskaraz mantentzen den elkarrizketari buruzkoak gazteleraz jaso behar dituztelako pazientea artatzen den momentuan bertan.

Proiektu honen bitartez, Osakidetzari egiten zaion ekarpena ez da bere politika linguistikoetan ezarritako helburuetara bakarrik mugatzen, jendartearen artean hazten doan eskaera bati ere aurre egiten baitzaio. Izan ere, osasun-langile eta pazienteen arteko komunikazioa hobetzeak arretaren kalitatea bera hobetzea dakar, bidean erabiltzailearen gogobetetasuna indartuz. Bestalde, pazienteek dokumentazioa euskaraz jasotzeko eskubidea izanik, itzulpena egiteko itzultzaile automatiko baten laguntza izateak itzulpenkostuak murrizten ditu.

Proiektuak hiru fase aurreikusten ditu:

1. Corpus elebidunaren bilketa, zeinean hala nahi duten osasun-langileek parte hartu ahal duten.

2. Itzultzaile automatikoaren garapena, aurreko corpus elebiduna zein bestelako corpus elebidun eta elebakarrak erabilita.

3. Itzultzaile automatikoaren ebaluazioa, non corpusaren bilketan parte hartu duten osasun-langileak gonbidatuko diren ebaluazioa egitera.

3.2 Txosten klinikoen sailkapena

Jasotako txostenak hainbat ezaugarriren arabera sailka daitezke, garatutako itzultzaile automatikoa entrenatzerakoan eta ebaluatzerakoan lagungarriak izan daitezkeenak. Gauzak horrela, corpusa biltzeko aplikazioa diseinatzerakoan ezaugarri hauei buruzko informazioa gordeta gera dadin bilatu da.

Lehenik eta behin, erabiltzaileei buruzko metadatuak daude. Horrela, corpusaren bilketan parte hartu nahi duten bolondresek web-aplikazioan kontu bat ireki behar dute, eta aplikaziora sartzen diren lehen aldian lan egiten duten osasun-erakundea zein den adierazi beharko dute.

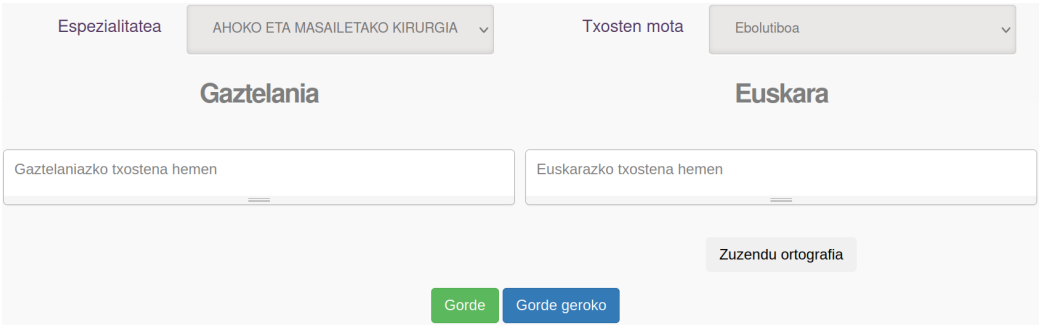
Osakidetza "Erakunde Sanitario Integratu" (ESI) deritzen unitateetan antolatua dago, pazientearen arretan ematen diren maila desberdinak integratzeko helburua dutenak, eta erakunde berean biltzen diren ospitale eta osasun-zentroak barnebiltzen dituztenak. Hauek geografikoki banatuta daude, eta Osakidetzaren erakunde guztien zerrenda osoarekin batera kontsulta daitezke.[1]

Bestalde, erabiltzaileek beren kontua sortzerakoan beraien lanpostua (erizaina, medikua, ...) adierazteko aukera dute, eta lan egiten duten espezialitatea edo arreta eremua adieraziko dute. Hala ere, osasun-langile batzuek espezialitate desberdinetan lan egin dezaketela kontuan hartuz, dokumentu bati dagokion espezializatea ez da langileari dagokiona, baizik eta dokumentuari berari esleitzen zaiona.

Beraz, dokumentu berri bat erregistratzerako orduan, dokumentuari dagokion espezialitatea (adib.: pediatria, larrialdiak, ...) eta txosten mota adieraziko dira. Honela, 3.1 irudiak corpusa biltzeko aplikazioa erakusten du, goiko partean espezialitatea eta txosten mota aukeratzeko menuak agertzen direlarik.

Espezialitateen zerrenda osoa 3.3. taulan kontsulta daiteke, eta txosten motei dagokienez hauek dira Osakidetzarekin batera definitu direnak:

[1] `https://www.osakidetza.euskadi.eus/gardentasuna-gobernu-ona/-/osakidetzako-zerbitzu-erakundeen-organigramak/`



3.1 irudia – Corpus elebiduna biltzeko aplikazioaren ikuspegi orokorra, goiko partean espezialitatea eta txosten mota aukeratzeko menuak agertzen direlarik.

1. **Ospitaleratze-txostenak**, zeinak paziente berri bat ospitalean onartzean betetzen diren.

2. **Txosten ebolutiboak**, pazientea ospitalean dagoen bitartean bere eboluzioa adierazteko idazten direnak.

3. **Alta-txostenak**, pazienteak ospitalean egonaldia amaitzean idazten direnak. Alta-txostenek gaixoaren egonaldiaren laburpena dute.

4. **Baimen informatuak**, pazienteek ebakuntza edo edozein arrisku mota suposatzen duen prozedura bati aurre egin behar dietenean erabiltzen direnak.

5. **Bestelakoak**, aurreko kategoriatan sartzen ez diren dokumentuak.

Osakidetzaren lehentasuna alta-txostenak eta ebolutiboak modu egokian itzultzea izanik, jasotako corpusetik ebaluazio-corpusa erauzterakoan bi txosten mota hauetako dokumentuak bakarrik kontuan hartu dira.

Azkenik, 3.1 irudiaren beheko partean ikus daitekeen moduan, corpusa biltzeko web-aplikazioan erabiltzaileak aukera dauka dokumentua behin-betirako gordetzeko edo beranduago editatzen jarraitzeko. Aldagai hau ere ebaluazio-corpusa erauzterakoan erabili da, ebaluaziorako erabiltzaileak amaitutzat emandako dokumentuetatik erauzitako esaldiak bakarrik erabili direlarik.

3.3 Corpusa biltzeko web-aplikazioa

Corpus egokia biltzeko beharrezkoa da erabiltzaileei gidalerro egokiak ematea. Hau horrela izanda, erabiltzaileei honako gidalerroak eskaini zitzaizkien txosten berri bat idazterakoan:

1. Jatorri eta helburuko testuak esaldi mailan parekatu, lerro bakoitzean esaldi bakarra idatziz. Jatorrizko esaldi batek helburu-hizkuntzan esaldi bat baino gehiago behar baditu, hauek lerro berean txertatu, eta berdin alderantzizko zentzuan.

2. Idazteko estilorik ahalik eta naturalena erabili. Corpus errealista behar dugu, beraz, erabilitako estiloa ere halakoa izan behar da.

3. Erregistro formala erabili; euskalkiei dagozkien aldaketa lexikoak onartzen dira, baina ez estandarizatutako terminoetatik aldentzen diren aldaketa ortografikoak.

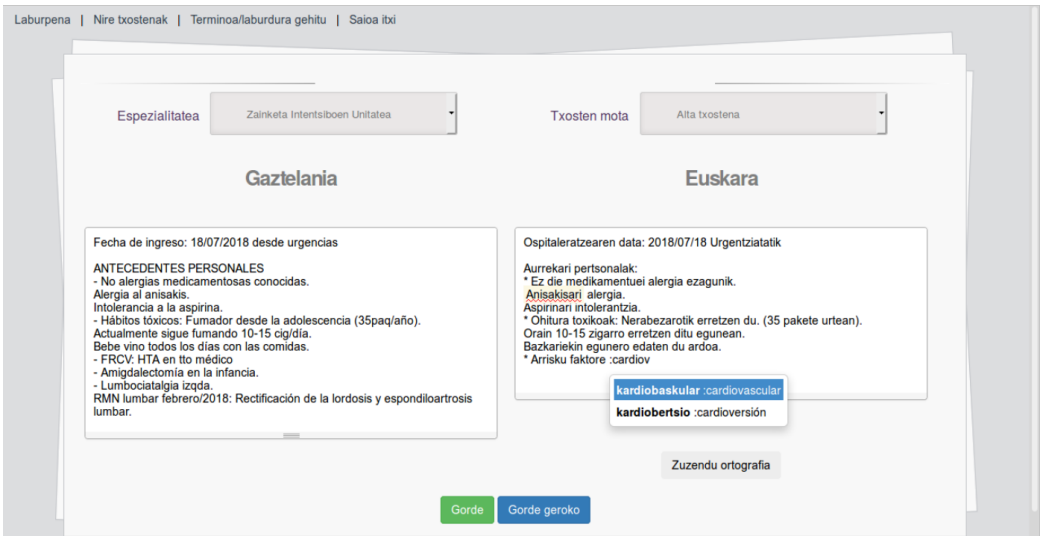
Gidalerro egokiak eskaintzeaz gain, garrantzitsua da osasun-langileen esfortzua ahal den heinean murriztea. Helburu horrekin, web-aplikazioan erabiltzaileentzako lagungarriak izan daitezkeen tresna batzuk integratu dira:

1. Donostia Unibertsitate Ospitalean euskaraz idatzitako ereduzko altatxostenak (Joanes Etxeberri Saria V. Edizioa, 2014).

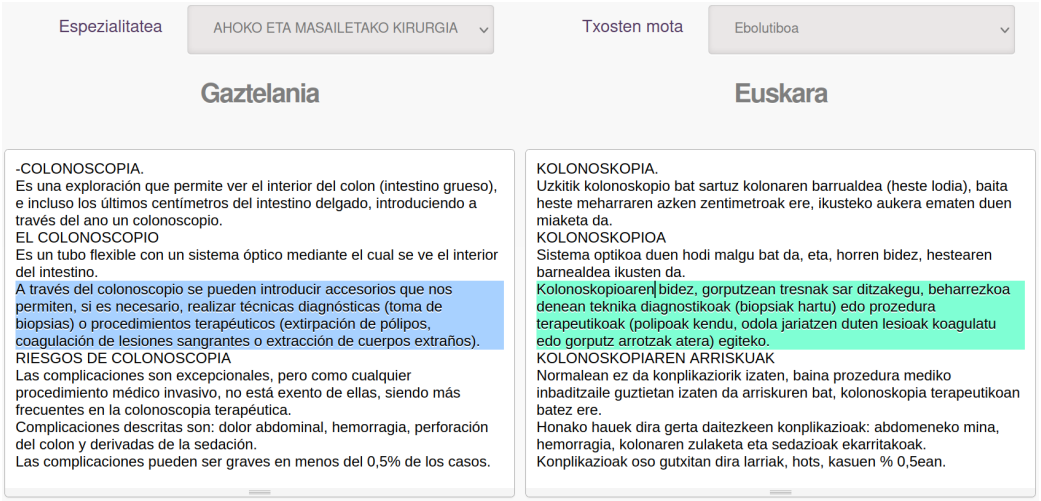
2. Hiztegi dinamiko bat, erabiltzaileek beraiek kudeatu dezaketena, osasunlangile batek termino bat idatzi baino lehen bere lankideek aurretik idatzitako proposamenak kontsultatzeko aukera emanez.

3. Euskarazko testua idazteko kutxan hiztegi kliniko integratu bat, ':' ondoren gaztelerazko terminoa idatziz euskarazko baliokidea eskaintzen duena. Hiztegi honen erabileraren adibide bat 3.2 irudian ikus daiteke.

Gainera, testu elebiduna esaldi mailan parekatua izan dadin bermatzen laguntzeko, bi tresna erabiltzen dira corpusa biltzeko web-aplikazioan: 1) testua editore batetik web-aplikaziora itsastean, automatikoki esaldika banatzea, lerro bakoitzean esaldi bana gordez; eta 2) momentu bakoitzean editatzen ari den esaldia kolorez markatzea, beste hizkuntzako esaldi baliokidea ere beste kolore batez azpimarratuz. Azken honi dagokionez, 3.3 irudiak hizkuntza bakoitzeko esaldiak nola azpimarratzen diren erakusten du.



3.2 irudia – Corpus elebiduna biltzeko aplikazioa, hiztegi laguntzailearen erabilera erakutsiz.



3.3 irudia – Corpus elebiduna biltzeko aplikazioa, hizkuntza bakoitzeko esaldiak azpimarratuta.

Azpiatal hau amaitu baino lehen, corpus bilketari eragiten dieten bi gai aipatu nahi ditugu, bilketa-prozesuaren antolaketari eta datuen pribatutasunari dagozkienak.

Antolakuntzari dagokionez, ESI bakoitzak proiektuaren arduradun bat dauka, zeinak corpusaren bilketan interesa izan dezaketen osasun-langile elebidunen zerrenda bat duen. Proiektura langile gehiago batzeko asmoz, COVID-19aren pandemia baino lehen hainbat bilera antolatu ziren, zeintzuetan Osakidetzako ordezkari instituzionalek, arduradun teknikoek eta proiektuan inplikatutako Ixakideek parte hartu zuten. Bilera hauetan proiektua aurkeztu, corpus bilketaren arlo teknikoak azaldu eta langileen zalantzak argitu ziren. Pandemia garaian, aurkezpen hauek langileen esku utzi zen bideo batez ordezkatu ziren. Bestalde, behin proiektua martxan jarrita, posta elektroniko bidez buletinak bidali zitzaizkien parte-hartzaileei. Buletin hauek bolondresak proiektuaren berri izan zezaten eta beraien esfortzuaren emaitzak ezagutu zitzaten bidaltzen ziren interesgarria erizten zenean.

Proiektuaren amaieran ESI bakoitzak erabakitzen du langileak egindako lanagatik konpentsatu ala ez, eta hala izatekotan, zein formatan (adibidez, egun libreak emanez).

Azkenik, jasotako txosten klinikoak ez dira errealak izan behar, baina hala izatekotan, ez dute pazienteari buruzko informazio pertsonalik eduki behar (adib.: izen-abizenak, jaiotze-data, ...). Horretarako, Osakidetzan pazienteari buruzko informazio pertsonala eta txosten klinikoak modu desberdinduan jasotzen dira, biak lotzeko kode bat erabiliz. Bestalde, behin itzultzaile automatikoa garatuta, hau Osakidetzako sistema informatikoan inplementatuko da, informazio jarioa barne-mailakoa izango dela bermatuz.

Horretaz gain, jasotako corpusarekin lan egin duten ikerlari guztiek konfidentzialtasun-konpromiso bat sinatu dute, eta sarreran aipatu bezala, corpus bakoitzeko dokumentu guztietatik erauzitako esaldiak ausaz berrordenatu dira, bertan jasotzen den informaziotik pazientea zein izan daitekeen jakitea ekiditeko asmoz. Pribatutasunari dagozkien bi neurri hauek, sakabanatzea eta desordenatzea, Itzulbideko corpus elebidunari eta corpus elebakar guztiei aplikatzen zaizkie. Corpusaren atzipena mugatua da eta corpusa ikerlari zehatz batzuek soilik atzi dezakete.

3.4 Corpus bilketaren emaitzak

Guztira, 210 bolondres erregistratu ziren Itzulbide aplikazioan corpusa biltzeko, horietatik 149k dokumentu elebidunen bat idatzi zutelarik.

Horien % 50,3k euskarazko 500 hitz baino gehiago idatzi zituzten; % 34,8k 1.000 hitz baino gehiago erregistratu zituzten; % 12,3k gutxienez 3.000 hitz

jaso zituzten; eta % 2,7k 10.000 hitzeko langa gainditu zuten.

Corpusa web-aplikaziotik CSV (jatorrizko ingelesez, *Comma-Separated Values*) batera esportatu zen, non dokumentu bakoitzeko euskarazko eta gaztelerazko esaldiez gain, aurretik aipatutako aldagai batzuk jasotzen ziren; besteak beste, erabiltzaileak dokumentua amaitutzat eman zuen ala ez adierazten duen balioa, eta dokumentuari dagozkion espezialitatea eta txosten mota.

Fitxategi honetatik dokumentu bakoitzeko esaldi pare elebidunak eta dagozkien aldagaiak eraucteko programa bat garatu zen, eta ondoren, eskuz errepasatu zen lortutako emaitza. Izan ere, nahiz eta lerro bakoitzean esaldi bakarra idazteko gidalerroa argia izan, eta corpusa jasotzeko web-aplikazioak horretan laguntzeko tresnak eduki, hainbat dokumentutan hori ez zen betetzen, lerrokatzeak eskuz aldatu behar izan zirelarik.

Horrela, jasotako CSVtik erauzitako corpusaren lehenengo bertsioak, hurrengo kapituluan *Itzulbide 1.0* deitu duguna, jarraian adierazten diren estatistikak ditu, guztira 26.437 esaldi pare elebidun izanda. Taula guztietan, dokumentu kopurua CSVtik ateratakoa da; esaldi eta token kopuruak erauzitako corpusari dagozkionak izanda.

Lehenik eta behin, 3.1. taulan estatistikak jatorrizko dokumentua amaitutzat eman zen ala ez adierazten duen aldagaiaren arabera erakusten dira, bakoitzari dagozkien dokumentu, esaldi eta token kopuruak adieraziz.

Egoera	**Dokum.**	**Esaldiak**	**Tok. (eu)**	**Tok. (es)**
Amaitua	1.774	23.695	198.503	236.462
Amaitu gabea	179	2.742	19.569	23.034

3.1 taula – Itzulbide 1.0 corpusaren estatistikak, dokumentuaren egoeraren arabera.

Ondoren, 3.2. taulan estatistikak txosten motaren arabera aurkezten dira, berriro ere bakoitzaren dokumentu, esaldi eta token kopuruak adieraziz.

Azkenik, 3.3. taulak espezialitate desberdinen distribuzioa erakusten du, bakoitzaren dokumentu, esaldi eta token kopuruekin.

Ikusten denez, espezialitate desberdinetako txostenen distribuzioa oso desorekatua da. Adibidez, 8 espezialitatetarako 100 dokumentu baino gehiago daude eskuragarri; beste 10 espezialitaterako 10 dokumentu baino gutxiago bildu diren bitartean. Espezialitate desberdinen artean, "familia-medikuntza" eta "larrialdiak" dira dokumentu gehien dituztenak.

Txosten mota	**Dokum.**	**Esaldiak**	**Tok. (eu)**	**Tok. (es)**
Ospitaleratze-txostenak	24	625	4.989	5.494
Txosten ebolutiboak	1.333	15.069	110.699	127.036
Alta-txostenak	260	4.424	29.660	33.174
Baimen informatuak	139	3.006	42.193	55.639
Bestelakoak	197	3.313	30.531	38.153

3.2 taula – Itzulbide 1.0 corpusaren estatistikak, txosten motaren arabera.

Corpus hau espezialitate eta txosten motaren arabera hainbat proba egiteko erabili zen, hortaz errepikatutako esaldiak ezabatzerakoan aldagai hauek ere kontuan hartu ziren. Hau da, gerta daiteke esaldi pare elebidun berdina espezialitate edo txosten mota desberdinarekin agertzea.

Azken sistema garatzeko, bigarren fase batean beste 7.046 esaldi pare elebidun jaso ziren, eta behin aurrekoei batuta, errepikatutako esaldiak ezabatu ziren; kasu honetan, espezialitatea eta txosten mota kontuan hartu gabe. Gainera, entrenamendu-corpusetik aurretik ebaluaziorako erauzitako corpusean agertzen ziren esaldiak baztertu ziren, hurrengo kapituluan *Itzulbide 2.0* deitu dugun azken corpusean guztira 30.805 esaldi pare elebidun bilduz.

Espezialitatea	**Dokum.**	**Esaldiak**	**Tok. (eu)**	**Tok. (es)**
Ahoko eta aurpegi-masailetako kirurgia	13	86	782	835
Ahoko eta masailetako kirurgia	1	22	323	453
Anestesia eta bizkortzea	22	124	812	947
Arnas aparatua	51	2.360	16.893	17.620
Barne-medikuntza	182	4.392	31.201	35.922
Digestio-aparatua	39	1.183	8.349	9.836
Egonaldi laburreko psikiatria	7	120	1.381	1.507
Emergentziak	4	2	9	36
Erizaintza	156	1.101	10.537	13.413
Erradiodiagnostikoa	40	240	1.738	2.135
Errehabilitazioa	10	53	321	338
Etengabeko arreta	51	413	3.827	4.276
Etxeko ospitalizazioa	76	642	4.077	4.685
Ezezaguna	104	2.413	29.594	38.624
Familia-medikuntza	251	2.483	17.198	19.023
Farmazia	50	795	8.534	11.024
Ginekologia eta obstetrizia	4	53	315	348
Kardiologia	1	25	213	280
Kirurgia orokorra	2	15	128	134
Kudeaketa sanitarioko unitatea	1	22	467	639
Larrialdiak	226	2.940	20.003	22.043
Medikuntza intentsiboa	33	643	4.811	5.674
Otorrinolaringologia	52	1.019	6.260	7.319
Pediatria	74	374	2.804	3.267
Prebentzio-medikuntza	1	23	108	123
Psikiatria	172	746	6.619	8.165
Traumatologia	104	1.219	9.054	10.950
Urologia	125	2.751	29.500	37.306
Zainketa aringarriak	97	156	1.915	2.278
Zainketa aringarrien unitatea	3	21	286	275
Zuzendaritza	1	1	13	21

3.3 taula – Itzulbide 1.0 corpusaren estatistikak, espezialitatearen arabera.

4. KAPITULUA

Baliabideak

Kapitulu hau bi ataletan banatuta dago: lehenengoan erabili ditugun corpusak deskribatuko ditugu, eta bigarrenean itzulpen automatikorako eta testuak aurreprozesatzeko erabili ditugun sistemak aipatuko ditugu.

4.1 Corpusak

Sarreran aipatu moduan, corpusetan oinarritutako itzultzaileei dagokienean, corpusa zenbat eta handiagoa eta domeinura egokituagoa izan, hainbat eta hobea izango da garatutako itzultzailearen kalitatea. Gure kasuan, ordea, Itzulbide proiektuan jasotako domeinu klinikoko corpus elebiduna txikia izanda, beharrezkoa izango da beste domeinuetako corpusak ere erabiltzea.

Edonola ere, kontuan izanda garatu nahi dugun itzultzailean zehaztasuna ahalik eta altuena izatea nahi dugula, beste domeinuetako corpusak biltzerakoan irizpide orokor batzuk jarraituko ditugu aukeratutako corpusen kalitatea ziurtatze aldera.

Horretaz gain, domeinu klinikoko itzultzailea garatzerakoan terminologia modu zuzenean itzultzea erronka nagusietako bat izanda, euskara eta gaztelera parerako eskuragarri dauden terminologia klinikoko baliabide guztiak erabiliko ditugu.

Bestalde, gaztelerazko testu kliniko kopuru handia eskuragarri izanda, hauek atzeranzko itzulpenaren bidez txertatu dira garatutako sistemetan. Esperimentu batzuetan, kopiatze teknikaren bidez ere gehitu ditugu corpus

elebakarretako batzuk; hau da, gaztelerazko corpusa euskarazkoa balitz bezala entrenamendu-corpusaren bi aldeetan txertatuz (Currey *et al.*, 2017).

Erabilitako corpusak deskribatzerakoan hauek 4 multzotan banatu ditugu: 1) domeinuz kanpoko corpus elebidunak, 2) terminologia kliniko eleanitza, 3) domeinu klinikoko corpus elebidunak, eta 4) domeinu klinikoko gaztelerazko corpusak.[1]

Baliabide hauek modu gradualean gehitu dira tesian zehar garatutako sistema desberdinetara; izan ere, corpus guztiak ez dira hasieratik eskuragarri egon. Honek aukera eman digu gehitutako corpus batzuek garatutako itzultzailearen kalitatean egindako ekarpena neurtzeko.

Orokorrean, diseinatutako esperimentuen arteko desberdintasun handiena ebaluaziorako erabilitako corpusean egon da. Itzulbideko corpus elebiduna jaso baino lehen, domeinu klinikoko corpus elebidunik ezean Donostia Unibertsitate Ospitalean bildutako euskarazko alta-txostenak (Joanes Etxeberri Saria V. Edizioa, 2014) eta haien gaztelerazko eskuzko itzulpenak erabili ditugu ebaluaziorako, nahiz eta hauek idaztearen helburua bestelakoa izan.

Jarraian, erabilitako corpusak eta beraien estatistikak aurkezten dira, aurretik definitutako multzoen arabera deskribatuta.

4.1.1 Domeinuz kanpoko corpus elebidunak

Gure atazarako erabilgarriak izan litezkeen domeinuz kanpoko corpus elebidunak hautatzerakoan bi helbururen arteko erdibidea bilatu dugu: 1) ahalik eta corpus gehien biltzea; eta 2) bildutako corpusen kalitatea minimo batetik gorakoa izatea, itzulpenen zehaztasuna ahalik eta altuena izan dadin.

Helburu horrekin, bi irizpide nagusi jarraitu ditugu erabili beharreko domeinuz kanpoko corpus elebidunak hautatzerakoan: 1) aurretik euskararen eta gazteleraren arteko itzultzaile automatikoak garatzeko erabili izana; edota 2) itzultzaile profesionalek itzuliak izana eta itzuli beharreko txosten klinikoen antzeko ezaugarriak izatea.

Jarraian, erabilitako domeinuz kanpoko corpus elebidunak zerrendatuko ditugu.

1. **EiTB (2016)**: Euskal Irrati TeleBistako albisteez osatutako corpusa,

[1] Aurrerago 4.1.2. atalean aipatzen dugun moduan, terminologia klinikoa zuzenean gehitu da entrenamendu-corpusera, horregatik hemen corpus bezala zerrendatzen dugu.

0,56M[2] esaldi pare elebiduneko kopuruarekin. Esaldi hauek ez dira paraleloak, konparagarriak baizik; hala ere, erabiltzea erabaki da beraien kalitatea frogatuta dagoelako (Etchegoyhen *et al.*, 2016). Corpus hau 3 aldiz errepikatuta erabili da esperimentu guztietan.

2. **HAEE**: Herri Arduralaritzaren Euskal Erakundearen testu administratiboak, guztira 0,9M esaldi pare elebidunekin.

3. **Consumer**: Eroskiren izen bereko aldizkaritik erauzitako esaldi pare elebidunak, kontsumo arloko 268.112 esaldi barnebilduz.

4. **Irrika**: Elhuyarrek gazteentzako prestatutako aldizkaritik erauzitako testuak, zientzia-dibulgazio arloko 5.570 esaldi pare elebidunez osatua.

5. **EIZIE**: Euskal Itzultzaile, Zuzentzaile eta Interpreteen Elkarteak bildutako itzulpen-memoriak, guztira 94.552 esaldi pare elebidunekin.

6. **Pelikulen sinopsiak**: filmen sinopsiez osatutako corpusa, 237.883 esaldi pare elebidun barnebilduz.

7. **PacoWebCorpus2012**: 82 webguneetatik PACO2 (San Vicente eta Manterola, 2012) tresna erabilita erauzitako 659.395 esaldi pare elebidun.

8. **HAC**: Hizkuntzen Arteko Corpusa (Sarasola *et al.*, 2015), literatura domeinuko 566.738 esaldi pare elebidunez osatua.

9. **Osakidetza profesionalak**: Osakidetzak bere langileentzat argitaratutako testuetatik erauzitako 22.051 esaldi pare elebidun. Esaldi hauek domeinuz kanpokotzat hartu dira beraien edukia itzuli nahi diren txosten klinikoekiko oso desberdina delako. Zehazki, Osakidetzako Profesionalentzako Argitalpenen ataleko webgunean[3] eskuragarri zeuden dokumentuak erabili ditugu, 'Planak eta programak' eta 'Memoriak' ataletako dokumentuak alde batera utzita.

[2]Corpusen tamainak aipatzerakoan, ahalik eta datu zehatzena ematea izan da jarraitutako irizpidea. Horregatik, esaldi kopuru zehatza eskuragarri dugunean kopuru hau adierazi dugu; eta ez daukagunean dezimalak erabili ditugu.

[3]`https://www.osakidetza.euskadi.eus/profesionalak/-/argitalpen-profesionalak/`

Dokumentu hauek PDF (jatorrizko ingelesez, *Portable Document Format*) formatutik testu hutsera pasa ziren, eta esaldiak banatzeko Itzulbideko corpusa eraucteko 3.4. azpiatalean aipatutako programa bera erabili zen. Azkenik, esaldien parekatzea eskuz egin zen, hizkuntza bateko esaldi bat beste hizkuntzan hainbat esalditan agertzean hauek ';' bidez bananduz.

10. **EiTB (2020)**: Aurreneko corpusaren bertsio berritua (Etchegoyhen eta Gete, 2020), EiTBko albisteez osatua. Guztira 637.182 esaldi pare elebidun ditu, eta errepikatu gabe gehitu da entrenamendu-corpusera.

Domeinuz kanpoko 1-7 corpusei dagokienez, hasiera batean modu gordinean gehitu baziren ere, momentu batetik aurrera hizkuntz-identifikatzaile bat[4] aplikatu ondoren gehitu ziren entrenamendu-corpusera. Modu horretan, maiztasun txikiko izen propioz edo orokorrean entitate izendunez osatutako esaldi asko kanpoan geratu ziren, domeinuz kanpoko hiztegia murriztuz eta hortaz, domeinu klinikoaren berezko terminologia itzultzeko beharrezkoa den lexikorako zati handiago bat utziz.

Hizkuntz-identifikatzailea aplikatu ondoren, 1-7 corpusei dagozkien esaldiak 3,7M izatera igaro ziren. Prozesu hau txosten klinikoen itzulpenaren kalitaterako onuragarria zela frogatu zen.

Bestalde, publikoki eskuragarri dauden bestelako corpus batzuk (adib.: 'OpenSubtitles' edo 'GNOME') erabiltzea baztertu dugu, itzulpenen kalitatea bermatua ez dagoelakoan.

Laburpen modura, 4.1. taulak domeinuz kanpoko corpus elebidunak eta haien domeinuak eta esaldi kopuruak erakusten ditu.

4.1.2 Terminologia kliniko eleanitza

Aurretik 2.2. azpiatalean aipatu bezala, erabilitako terminologia klinikoetatik handiena SNOMED CTren euskaratzetik dator (Perez-de Viñaspre, 2017). Prozesu hau tesi honen garapenarekiko paraleloa izanda, SNOMED CTtik eratorritako terminologiaren bi bertsio erabili ditugu tesian zehar.

Horretaz gain, GNS-10aren euskaratzea ere baliatu dugu gaztelerazko termino baliokideekin batera entrenamendu-corpusera gehitzeko. Corpus honek 2 bertsio ditu, 2020. eta 2021. urteetan WMT konferentzian biomedikuntzaren arloko testuak itzultzeko atazan eskuragarri utzitakoak.

[4] `https://github.com/saffsd/langid.py`

Corpusa	Domeinua	Esaldi kopurua
EiTB (2016)	Albisteak	0,56M (x3)
HAEE	Administratiboa	0,9M
Consumer	Kontsumoa	268.112
Irrika	Zientzia-dibulgazioa	5.570
EIZIE	Itzulpen-memoriak	94.552
Pelikulen_sinopsiak	Filmen sinopsiak	237.883
PacoWebCorpus2012	*Web-crawling*	659.395
HAC	Literatura	566.738
Osakidetza_profesionalak	Osasuna/administratiboa	22.051
EiTB (2020)	Albisteak	637.182

4.1 taula – Domeinuz kanpoko corpus elebidunak eta haien domeinuak eta esaldi kopuruak

Azkenik, COVID-19arekin lotutako termino gutxi batzuk ere bildu ditugu, gure sistemak sortutako terminologia berrienetara egokitzeko asmoz.

Terminologia hauek guztiak ingeleserako ere eskuragarri daudenez, WMT 2020ko Biomedical atazarako ingelesezko sarrerak ere erabili ditugu.

Edonola ere, garatutako sistema guztietan terminologia hauek modu zuzenean txertatu dira; hau da, bestelako esaldiak balira bezala. Itzuli nahi ditugun txosten klinikoetan esaldiak laburrak izan ohi direnez, espero da honek kalterik ez eragitea.

Sistemak zuzenean elikatzeko erabiltzeaz gain, hasierako esperimentu batzuetan SNOMED CTko erlazioetan oinarrituta esaldi artifizialak sortu dira; eta egindako errore analisi batean terminologia klinikoaren itzulpena ebaluatzeko ere erabili dira.

Jarraian, erabilitako terminologia kliniko bakoitza deskribatuko da:

1. **SNOMED CT 1.0** SNOMED CTren euskaratzetik eratorritako corpusaren 1. bertsioa, gaztelerazko 83.360 terminoei dagozkien 151.111 gaztelera-euskara termino pareekin (gaztelerazko termino batek euskarazko ordain bat baino gehiago izan ditzake). Espero da gaztelerazko termino bakoitzarentzako euskarazko sarrera bat baino gehiago egotea lagungarria izatea euskaratik gaztelerara itzultzeko.

2. **SNOMED CT 2.0** SNOMED CTren euskaratzetik eratorritako corpusaren 2. bertsioa, 11 tokeneko luzerararteko terminoak jasoz, guztira 896.898 termino elebidun izanik.

3. **GNS-10 1.0**: Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioan jasotako terminoak euskaratzetik eratorritako corpusa, 2020ko bertsio honetan 27.696 termino elebidun dituena.[5]

4. **GNS-10 2.0**: Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioan jasotako terminoak euskaratzetik eratorritako corpusa, 2021eko bertsio honetan 29.670 termino elebidun dituena.[6]

5. **SNOMED CTren COVID-19arekin lotutako barne-argitalpena**: 2020ko martxoan SNOMED CTk argitaratutako 84 terminoen bilduma,[7] Osakidetzako itzultzaile profesional batek euskaratua.

6. **Elhuyarrek argitaratutako COVID-19arekin lotutako terminoak**: 2020ko lehen seihilekoan zehar Elhuyarrek bere webgunean eskuragarri utzitako 126 terminoen bilduma. Termino hauek tesi honetarako prestatu ondoren jada ez daude webgunean bilduta.

Azpiatal hau amaitzeko, 4.2. taulak erabilitako terminologia klinikoaren esaldi eta token kopuruak adierazten ditu.

Terminologia	**Terminoak**	**Tokenak (eu)**	**Tokenak (es)**
SNOMED CT 1.0	151.111	271.244	257.639
SNOMED CT 2.0	896.898	3.074.750	5.309.227
GNS-10 1.0	27.696	229.248	175.627
GNS-10 2.0	29.670	245.150	188.233
SNOMED CT / COVID-19	84	579	729
Elhuyar / COVID-19	126	263	243

4.2 taula – Terminologia klinikoak, beren termino eta token kopuruekin

4.1.3 Domeinu klinikoko corpus elebidunak

Azpiatal honetan 3 corpus deskribatuko ditugu, azkenekoak bi bertsio dituelarik:

[5] `https://drive.google.com/drive/folders/1ooKi1sDF-nne0DxrepzQZMwSALG8YCDS`

[6] `https://drive.google.com/drive/folders/1gUQHoutvYIXGGPVTBbBF3qlHHhX9qbrO`

[7] `http://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release%2DCOVID-19`

1. **Donostia Unibertsitate Ospitaleko alta-txosten ereduak** (Joanes Etxeberri Saria V. Edizioa, 2014), espezialitate anitzetako 42 dokumentuz osatua, Osakidetzako mediku elebidun batek gaztelerara itzuliak. Tesi honen lehen erdian sistemak ebaluatzeko baliagarriak izan ziren, 1.038 esaldi garapenerako eta beste 1.038 probarako erabili zirelarik. Jatorrizko dokumentu hauek helburu akademikoekin idatziak izanda, beraien ezaugarriak benetako txosten klinikoekiko desberdinak dira, idazterakoan hizkuntzaren zuzentasunari lehentasuna ematen baitzaio. Eskuzko itzulpena mediku batek egindakoa izanda, ordea, gaztelerazko esaldien ezaugarriak benetako txostenekiko antzekoagoak dira, orokorrean esaldi laburragoak erabiliz, ahal denean laburdurak txertatuz, eta euskarazko esaldi baliokideekin alderatuta akats ortografiko gehiago dituztelarik. Hau guztia kontuan hartuta, behin Itzulbideko corpus elebiduna eskuragarri izan genuenean erabiltzeari utzi zitzaien.

2. **Basurtoko Unibertsitate Ospitaleko saio klinikoak**: E3C proiekturako bildutako saio kliniko eleanitzen artean (Magnini *et al.*, 2020), Basurtoko Unibertsitate Ospitalean euskaraz eta gazteleraz idatzitako 17 saio klinikoen bilduma, guztira 541 esaldiz osatua. Testu hauek sarean eskuragarri daude,[8] eta Itzulbideko corpus elebidunarekin batera sistemak birdoitzeko erabili ziren.

3. **Itzulbide 1.0**: Itzulbide proiektuaren baitan 2020ko apirilaren 10era arte jasotako esaldi pare elebidunak. 26.437 esaldietatik 1.000 garapenerako eta beste 1.000 probarako gorde ziren. Corpus honekin egindako esperimentu batzuetan esaldi bakoitzaren espezialitate eta txosten mota kontuan hartu ziren; hortaz, esaldi pare berdina espezialitate edo txosten mota desberdinetan agertzen bazen, sarrera desberdin bezala kontsideratzen zen errepikatutako esaldiak baztertzerakoan.

4. **Itzulbide 2.0**: Itzulbide proiektuaren amaieran, 2021eko apirilaren 12rarte jasotako corpus elebiduna. Bertsio honetatik, entrenamendurako 28.805 esaldi erabili ziren, eta garapen eta probarako aurreko bertsioko esaldi berdinak mantendu ziren. Kasu honetan, azken sistemetan erabili beharreko corpusa izanik, esaldi errepikatu guztiak alde batera utzi ziren, nahiz eta espezialitatea edo txosten mota desberdina izan.

[8] `https://github.com/hltfbk/E3C-Corpus`

Azkenik, 4.3. taulak erabili ditugun domeinu klinikoko corpus elebidunak zerrendatzen ditu, bakoitzaren esaldi eta token kopuruak adieraziz.

Corpusa	**Esaldiak**	**Tokenak (eu)**	**Tokenak (es)**
Donostia Unibertsitate Ospitalea	2.076	19.938	19.022
Basurtoko Unibertsitate Ospitalea	541	5.254	5.185
Itzulbide 1.0	26.437	218.072	259.496
Itzulbide 2.0	30.805	353.986	392.607

4.3 taula – Domeinu klinikoko corpus elebidunak, beren esaldi eta token kopuruekin

4.1.4 Domeinu klinikoko gaztelerazko corpusak

Osakidetzak eskuragarri utzitako dokumentu kliniko gehienak gaztelera hutsean idatziak daude. Lehen esperimentuetan Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko txosten klinikoak erabili ditugu, eta azken sistemetan integratzeko Itzulbide proiektuan jasotako corpus handiago batzuk erabili ditugu, bestelako corpusen tamainei egokitzeko datuen hautespena aplikatu dugularik.

Jarraian, erabili ditugun domeinu klinikoko gaztelerazko corpusak deskribatuko dira. Dokumentu hauek atzeranzko itzulpena egiteko erabili direnez, itzuli baino lehen errepikatutako esaldiak ezabatu dira. Beraz, corpus bakoitzaren estatistikak aipatzerakoan jatorrizko esaldi kopuruaz gain, errepikatu gabeko esaldien kopurua ere adierazten dugu.

1. **Galdakao-Usansolo Ospitaleko alta-txostenak**: 142.154 dokumentuz osatutako corpusa, jatorrizko 4.363.627 esaldietatik errepikatu gabeko 2.023.811 esaldi atzeranzko itzulpena egiteko erabili direlarik. Corpus hau egindako esperimentu guztietan erabili zen. Momentu batetik aurrera bertsio murriztu bat erabili zen, dokumentuaren identifikatzailea edo data soilik duten lerroak ezabatu ondoren 1.921.672 esaldirekin geratuz.

2. **Basurtoko Unibertsitate Ospitaleko alta-txostenak**: 57.569 dokumentu barnebiltzen dituen corpusa, jatorrizko 2.713.424 esaldietatik

dokumentu bakoitzaren lehen 2 lerroetan agertzen ziren kodeak eta datak kenduta, eta errepikatutako esaldiak alde batera utzita 905.893 esaldi erabilgarri geratuz.

3. **Basurtoko Unibertsitate Ospitaleko txosten ebolutiboak**: 471.500 dokumentuz osatutako corpusa, guztira 4.811.294 esaldi dituena. Dokumentu bakoitzaren lehen lerroan agertzen ziren kodeak eta datak ezabatuta, eta errepikatutako esaldiak kenduta, 2.318.361 esaldi kopuruan geratu zen.

4. **Itzulbide proiektuko corpus elebakarra**: corpus hau 4 zatitan banatuta dago: 1) Txosten ebolutiboetatik erauzitako 49.069.600 esaldi, 2) Traumatologia espezialitateko 2.412.202 esaldi[9], 3) Ospitaleratze-txostenetatik ateratako 6.550.241 esaldi, eta 4) Larrialdietako txostenetatik erauzitako 18.576.314 esaldi. Corpus hauen tamainak kontuan hartuta, eta Osakidetzaren lehentasuna alta-txostenak eta txosten ebolutiboak itzultzea izanda, lehenengo bi azpimultzoak bakarrik erabili dira, esaldi hutsak eta errepikatutakoak kenduta txosten ebolutiboetatik ateratako 18.667.813 esaldi eta traumatologiako 1.010.420 esaldi edukita. Datuen hautespena aplikatu ondoren, garatutako azken sistemetan erabili ziren corpus hauek.

   Domeinuz kanpoko corpusa 5M esaldi ingurukoa izanda, eta Galdakao-Usansolo Ospitalekoak eta Basurtoko Unibertsitate Ospitalekoak batuta ere 5M esaldi inguru izanda, Itzulbideko corpus elebakarretik beste 5M esaldi aukeratu dira datuen hautespena aplikatuz. Modu honetan, azken sistemetan erabilitako corpusean atzeranzko itzulpenaren bidez erabilitako corpusa, corpus elebidunaren tamainaren bikoitza baino askoz handiagoa ez izatea bermatzen da, itzultzailearen kalitatea arriskuan jarri gabe (Poncelas *et al.*, 2018b).

Amaitzeko, 4.4. taulak erabili ditugun domeinu klinikoko gaztelerazko corpusak zerrendatzen ditu, bakoitzaren dokumentu mota zein esaldi eta token kopuruak adieraziz. Corpus guztien kasuan, jatorrizko esaldi eta token kopuruak adierazten ditugu (esaldi hutsak, errepikatutakoak edo kodeak eta datak bakarrik dituzten esaldiak kendu gabe).

[9]Espezialitate honetako txostenak ebaluazio berezitua egiteko asmoarekin banatu ziren

Corpusa	Dokumentu mota	Esaldiak	Tokenak (es)
Galdakao-Usansolo Ospitalea	Alta-txostenak	4.363.627	47.417.680
Basurtoko Unibertsitate Ospitalea	Alta-txostenak	2.713.424	17.144.473
Basurtoko Unibertsitate Ospitalea	Txosten ebolutiboak	4.811.294	29.047.905
Itzulbide (es)	Txosten ebolutiboak	49.069.600	270.753.011
Itzulbide (es)	Traumatologia esp.	2.412.202	12.521.055
Itzulbide (es)	Ospitaleratze-txostenak	6.550.241	43.761.415
Itzulbide (es)	Larrialdi-txostenak	18.576.314	97.488.753

4.4 taula – Domeinu klinikoko gaztelerazko corpusak, beren dokumentu mota, esaldi eta token kopuruekin

Azkenik, ikuspegi orokor bat izateko, 4.5. taulak eskuragarri dauden corpus mota desberdinen esaldi kopuruak aurkezten ditu, terminologia klinikoaren kasuan terminoak esaldi bezala zenbatuz. Bertsio bat baino gehiago dituzten corpusen kasuan, azkeneko bertsioa kontuan hartu da; eta corpus elebakarren kasuan, jatorrizko esaldi kopuruak batu dira. Aurrerago, 5. eta 6. kapituluetan esperimentu bakoitzean erabilitako corpus zehatzak aipatuko dira.

Corpus mota	Esaldiak
Domeinuz kanpoko corpus elebidunak	5.071.483
Terminologia kliniko eleanitza	926.778
Domeinu klinikoko corpus elebidunak	33.422
Domeinu klinikoko gaztelerazko corpusak	88.496.702

4.5 taula – Corpus mota desberdinen esaldi kopuruak

4.2 Sistemak

Atal honetan erabili ditugun itzultzaile automatikoak zerrendatuko ditugu, teknika desberdinen arabera antolatuz (EOIA, IAE eta IAN).

Bestalde, atalaren amaieran corpusak aurreprozesatzeko erabili ditugun tresnak aipatuko ditugu.

4.2.1 Erregeletan Oinarritutako Itzulpen Automatikoa

Euskararen eta gazteleraren artean EOIA bidez itzulpenak egiteko SNOMED CTren euskaratzerako garatutako *MatxinMed* tresna erabili dugu; hau da, *Matxin* (Mayor, 2007) medikuntza-domeinura egokitua, bere lexikoietan hiztegi klinikoak txertatuz.

Egindako esperimentuak alemanaren eta ingelesaren arteko itzulpenera hedatzerakoan, *Apertium* (Forcada *et al.*, 2011) tresna erabili dugu, mundu mailan EOIA egiteko erreferentziazko software irekia.

Bi hizkuntza pareetan, EOIA atzeranzko itzulpena egiteko erabili da; hau da, corpus elebakarrak gazteleratik euskarara eta ingelesetik alemanera itzultzeko.

EOIA erabili zen lehenengo esperimentu multzoan, euskararen eta gazteleraren artean atzeranzko itzulpena egiteko teknika eta arkitektura desberdinak konparatu ziren; bigarrenean, ordea, teknika desberdinekin atzeranzko itzulpena egin ondoren datuen hautespena aplikatu zen, kasu honetan esperimentuak goian aipatutako bi hizkuntza pareetarako egin zirelarik.

4.2.2 Itzulpen Automatiko Estatistikoa

Kasu honetan tresna bakarra erabili zen IAE hizkuntza pare desberdinetan egiteko, modu ia esklusiboan erabiltzen den *Moses* (Koehn *et al.*, 2007) tresnaren bidez.

Zehazki, defektuzko balioekin erabili zen *Moses*, hitzen alineaziorako *MGIZA* (Och eta Ney, 2003) erabiliz, berrordenatze-eredu bezala lexikalizatutako '*msd-bidirectional-fe*' eredua erabiliz, eta hizkuntza-eredurako *KenLM* (Heafield, 2011) erabiliz, hitz-segidaren tamaina 5 izanda. Itzulpen-eredua doitzeko *MERT* (jatorrizko ingelesez, *Minimum Error Rate Training*) (Och, 2003) tresna erabili zen, gordetako aukera hoberenen zerrendaren tamaina 100 zelarik.

EOIAren kasuan bezala, IAE atzeranzko itzulpena egiteko erabili da; euskararen eta gazteleraren artean teknika eta arkitektura desberdinak konparatzeko, eta aurreko hizkuntza pareaz gain, alemanaren eta ingelesaren artean teknika desberdinekin egindako atzeranzko itzulpenean datuen hautespena aplikatzeko.

4.2.3 Itzulpen Automatiko Neuronala

IANaren barruan bi arkitektura erabili genituen: Sare Neuronal Errekurrenteak (SNE) eta Transformer. SNEen kasuan tamaina desberdinetako GRU eta LSTM unitate motak probatu ziren.

Tesian egindako lehenengo esperimentuetan *Nematus* (Sennrich *et al.*, 2017) tresna erabili zen, garaian artearen egoera zena. Zehazki, hiperparametroen balio desberdinak probatzeko, SNOMED CT terminologia klinikoaren lehenengo bertsioa txertatzeko eta Galdakao-Usansolo Ospitaleko corpus elebakarra gehitzeko erabili zen lehen aldiz.

Aurrerago, atzeranzko itzulpena egiteko teknika eta arkitektura desberdinak probatzeko ere erabili zen, kasu honetan, besteak beste, geruza kopuru desberdinetako bi SNE konparatuz.

Behin *Nematus* tresna eguneratzen jarraituko ez zutela publiko egin zenean, kode irekiko *OpenNMT* (Klein *et al.*, 2017) tresna erabiltzen hasi zen, esperimentu gehienetan Transformer arkitekturaren PyTorch inplementazioa erabiliz. Inplementazio hau oso erabilterraza da, eta GPUen memoriaren erabilera eraginkorra egiten du, momentu oro erabilgarri dagoen memoria kapazitate ia osoa erabiliz.

Transformer arkitektura erabiltzean defektuzko hiperparametroak erabili ziren;[10] eta esperimentu batzuetarako LSTM arkitektura ere erabili zen, kasu honetan 4 geruza, geruza bakoitzeko 512 neurona, 0,2ko *dropout* balioa eta 128ko *batch* tamainarekin.

Azkenik, tesiaren amaierako fasean *Fairseq* (Ott *et al.*, 2019) tresna erabili zen; garaiko artearen egoera izateagatik, eta *OpenNMT*-rekin entrenatutako sistemak birdoitzeko zeuden arazoak ekiditeko.

Kasu honetan, Transformer arkitektura erabili zen, entrenamendu egonkorra lortzeko blog sarrera honen[11] amaieran adierazten diren hiperparametroak erabiliz.

Fairseq-en beste abantailetako bat entrenamendua azkartzeko itzulpenetan kalitaterik galdu gabe sare neuronaletan beharrezkoak diren kalkuluen doitasuna murrizteko *'fp16'* aukera izatea da, modu honetan tamaina bereko Transformer sistemak denbora murritzagoan entrenatzeko aukera emanez, eta beraz, kontsumitutako energia murriztuz.

Kontrara, *OpenNMT*-rekin alderatuta, *Fairseq*-ek GPUen memoriaren

[10] `http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model`

[11] `http://cslab.org/blog/fairseq-basics`

erabilera azpi-optimoa egiten du, entrenamenduaren momentu askotan memoria erabilera % 100etik urrun egonda. Honen ondorioz, entrenamenduak behar baino luzeagoak izateaz gain, kontsumitutako energia estimatzeko egindako kalkuluen zehaztasuna baxuagoa izango da.

4.2.4 Aurreprozesua

Atal hau amaitzeko, esperimentu guztietan aplikatutako aurreprozesua deskribatuko dugu.

Corpusetan oinarritutako (IAE eta IAN) itzultzaileak entrenatzeko testua *Moses*-en eskuragarri dauden tokenizatzaile[12] eta *Truecaser*[13] tresnen bidez prozesatu ditugu.

Truecase prozesuaren kasuan, 4.1. atalean deskribatzen diren domeinuz kanpoko 1-7 corpusak *Truecase* formatuan bakarrik edukita, gainontzeko corpusei aplikatu zaien eredua corpus hauetan ikasitakoa izan da esperimentu guztietan. Espero da honek itzulpenen kalitatean duen eragina baxua izatea.

Bestalde, IAN teknika erabiltzerakoan hitzen segmentazioa aplikatu da, BPE teknikaren jatorrizko inplementazio erabiliz.[14] Esperimentu guztietan BPE eredua entrenamendu-corpus osoan ikasi da; eta azken esperimentuetan, egileek gomendatu bezala hiztegi desberdinak gorde dira hizkuntza bakoitzerako, hizkuntza bakoitzean agertzen diren azpi-hitzak hizkuntza horretan bakarrik aplikatuz. Sistema batzuk entrenatzeko BPE-dropout (Provilkov *et al.*, 2020) erregularizazioa aplikatu da 0,1eko probabilitatearekin. Honek entrenamendurako beharrezkoa den denbora asko luzatzen du, egin beharreko *epoch* bakoitzeko entrenamendu-corpusaren kopia bat sortu behar delako, baina txosten klinikoetan ohikoak diren akatsen eragina murrizteko lagungarria izango delakoan aplikatu dugu. Esperimentu guztietan, euskararen eta gazteleraren artean itzultzerakoan hiztegiaren tamaina 90.000 izan da; WMT 2020ko atazan ingelesaren eta gazteleraren artean itzultzeko 32.000ko balioa erabili da; eta alemanaren eta ingelesaren artean itzultzerakoan, 89.500 lehen esperimentu multzoan eta 40.000 bigarrenean.

Azkenik, egindako esperimentu batzuetan *Moses*-en eskuragarri dagoen

[12] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl`

[13] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl`

[14] `https://github.com/rsennrich/subword-nmt`

corpusak garbitzeko tresna[15] erabili da. Adibidez, *Fairseq*-ekin atzeranzko itzulpenaren bidez corpus elebakarrak itzultzerakoan defektuz ez dira sistemak onartzen dituen baino esaldi luzeagoak ezabatzen; hortaz, memoria-erroreak ekiditeko corpus elebakarretako esaldien tamaina 1024 token baino gutxiagokoetara murriztu da.

[15] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl`

5. KAPITULUA

Metodologia eta emaitzak: domeinuko corpus elebidunik gabe

Kapitulu honetan eta hurrengoan tesian zehar egindako esperimentuak azaldu eta beraien emaitzak aurkeztuko ditugu. Horretarako, aztertutako aldagai bakoitzari atal bat eskainiko diogu, bakoitzean landutako helburu nagusia deskribatu, erabilitako corpusak eta sistemak aipatu, eta ateratako ondorio nagusiak zerrendatuz. Errore analisia edo giza ebaluazioa egin den kasuetan, hauek atalaren amaieran kokatu dira.

Egindako esperimentuak bi bloke nagusitan banatzen dira, baldin eta domeinuko corpus elebiduna eskuragarri genuen ala ez. Domeinuko corpus elebidunik gabeko egoeran, ebaluaziorako alta-txosten ereduak eta haien eskuzko itzulpenak erabili ziren. Behin Itzulbide proiektuko txosten elebidunak eskura izanda, hauek sistemak entrenatzeko zein ebaluatzeko erabili ziren. Kapitulu honetan ebaluaziorako alta-txosten ereduak erabiliz egindako esperimentuak azalduko dira. Esan beharrik ez dago, ebaluazio-corpus desberdinekin lortutako emaitzak ez dira bata bestearekin konparagarriak. Hala ere, ebaluazio-corpusa aldatzerakoan momentuan eskuragarri zegoen entrenamendu-corpus berdinarekin bi ebaluazio-corpus desberdinetan lortutako emaitzak aurkeztuko ditugu, bakoitzarekin lortutako emaitzak hobeto interpretatzen laguntzeko asmoz.

Kapitulu honetan, tesiaren lehen esperimentuetan itzultzailea garatzeko orduan izandako hainbat erronkei aurre egin nahi izan genien, domeinuko corpus elebidunaren gabeziak baldintzatuta. Hasiera moduan, garaian artearen

egoera ziren SNEen hiperparametroak momentuan genituen domeinuz kanpoko corpusekin emaitza optimoak lortzeko moldatu genituen (5.1. atala).

Behin hiperparametro optimoak izanda, domeinura egokitzeko baliabideak banaka gehitu ziren, bakoitzaren ekarpena neurtzeko eta ondorioak hurrengo esperimentuetan kontuan hartzeko asmoz (5.2. atala). Atal honetan txertatzen da SNOMED CTko erlazioetatik abiatuta sortutako esaldi artifizialekin egindako proba.

Ondoren, IAN sistema desberdinak probatu ziren (5.3. atala), geruza kopuru desberdinetako SNEetatik garaian artearen egoera berria bihurtu zen Transformerreraino. Honekin batera, atzeranzko itzulpena egiteko EOIA eta IAE teknikak ere erabili ziren, domeinuko corpusik gabeko eszenatoki batean lagungarri izan litezkeen aztertzeko.

Aurrerago, atzeranzko itzulpena egiteko teknika desberdinak alderatzeaz gain, haien bidez sortutako corpus sintetikoak batu, bakoitzaren aniztasun lexikala aztertu, eta datuen hautespenaren bidez emaitzak hobetzeko saiakerak egin ziren (5.4. atala). Kasu honetan, euskararen eta gazteleraren artean egindako esperimentu berdinak alemanaren eta ingelesaren arteko itzulpenean errepikatu ziren.

Azkenik, momentuko baliabideekin WMT2020ko biomedikuntzaren arloko testuen itzulpenerako ataza partekaturako sortutako sistemak aurkeztuko dira (5.5. atala). Kasu honetan helburua ingelesetik euskarara itzultzea zen, eta atazan ohikoa den moduan artikulu zientifikoen laburpenak itzultzeaz gain, terminologia klinikoa itzultzeko ere erabili ziren garatutako sistemak.

Jarraian, urrats hauek guztiak azalduko ditugu atalez atal.

5.1 Oinarrizko IAN sistemetan hiperparametroak optimizatzea

Helburua

IAN sistemen hiperparametroak domeinuz kanpo optimizatzea, itzultzaileen kalitatea hobetzeaz gain, ondoren egin beharreko domeinuari egokitzea sistema optimizatu baten gainean egiteko helburuarekin.

Corpusak

Garatu beharreko sistemen abiapuntu moduan, tesi hau hasteko momentuan *MODELA* proiektuaren barruan garatu berri ziren euskararen eta gazteleraren arteko IAN sistemak (Etchegoyhen *et al.*, 2018) hartu ziren eredu moduan. Horretarako, sistema hauek garatzeko erabilitako corpusaren antzeko bertsio bat erabili zen gure lehenengo sistemak garatzeko. Zehazki, 4.1.1. azpiatalean aipatutako domeinuz kanpoko 1-7 corpusak erabili ziren sistemak entrenatzeko, guztira 4.530.683 esaldirekin. Esperimentuen helburua hiperparametroak optimizatzea izanik, garapen eta proba multzoak entrenamendu-corpusaren jatorri berdinetik erauzi ziren. Garapen multzoak 3.482 esaldi zituen, eta proba multzoak 3.405.

Esperimentuak

Lehenengo esperimentu hauetan, garaian artearen egoera ziren SNEak erabili ziren arkitektura moduan, eta sare neuronalen hainbat hiperparametroen balio desberdinak probatu ziren emaitzak hobetzeko asmoz. *Nematus* (Sennrich *et al.*, 2017) tresna erabili zen, eta oinarri-lerro moduan Etchegoyhen *et al.* (2018) lanean erabilitako hiperparametroak erabili ziren. Sistema hau bi noranzko SNEa zen, geruza bakarreko 1.024 neuronekin; eta garatutako sistema guztietan bezala, aurreprozesuan 90.000 iterazioz ikasi zen BPE eredua. Bestalde, ikasketa-tasa 0,0001 izan zen eta *drop-out* ez zen erabili. Dekodetze-prozesuan, *length-normalization* eta *coverage-penalty* hiperparametroek 1 eta 0 balioak hartu zituzten, hurrenez hurren.

Aurreko balioak finkoak izanda, beste hiperparametro batzuen balio desberdinak probatu ziren, gehienbat Britz *et al.* (2017) lanean lortutako emaitzetan oinarrituta. Proba hauek bi noranzkoetan egin ziren (eu-es eta es-eu), eta modu sekuentzialean aplikatu ziren; hau da, hiperparametro baten balio desberdinak probatu ziren, eta hurrengo esperimentuetarako norabide bakoitzean emaitza hoberenak lortzen zituzten balioak erabili ziren. Probatutako balioak 5.1. taulan aurkezten dira, esperimentuak egiteko erabili zen orden berdinean, lerro bakoitzean ezkerrean agertzen den balioa *MODELA* proiektuan erabili zena izanik. Esperimentu hauen helburua, sistemak berak hobetzeaz gain, ondoren egin behar zen domeinuari egokitzea sistema optimizatu baten gainean egitea izan zen, emaitzak esanguratsuagoak izan zitezen.

Hiperparametroa	**Balioak**
Optimizatzailea	Adadelta / Adam
Unitate mota	GRU / LSTM
Beam-width	6 / 10
Esaldi multzoaren tamaina	30 / 32 / 64
Hitz-bektorearen tamaina	500 / 512 / 1024

5.1 taula – Sistemak domeinuz kanpo optimizatzeko probatutako hiperparametroen balioak

Emaitzak

Jarraian, 5.2. taulan hiperparametro bakoitzaren balio desberdinekin garapen (*dev*) eta proba (*test*) multzoetan lortutako BLEU balioak aurkezten dira. Hiperparametro bakoitzarentzat, hurrengo esperimentuetarako erabili zen balioari dagokion lerroa hitz lodiz adierazten da. Kasu gehienetan hiperparametro bakoitzarentzat proba multzoan BLEU altuena lortzen zuen balioa aukeratu zen, baina desberdintasuna txikia zenean garapen multzoko emaitzak ere kontuan hartu ziren. Kasu bakan batean, esaldi multzoaren tamaina 64ra igotzean, balio hau lehenetsi zen nahiz eta eu-es zentzuan 32ko balioarekin emaitza hobeak lortu garapen eta proba multzoetan, desberdintasun hauek txikiak zirelako eta kontrako zentzuan aukeratutako balioarekin koherenteak izateko. Bestalde, eu-es zentzuan 1024ko hitz-bektorearen tamainarekin esperimentua ezin izan zen burutu, memoria-errore batengatik.

Ondorioak

Ikus daitekeenez, hiperparametro desberdinak aldatzerakoan lortutako emaitzak orotarikoak izan ziren. Kasu batzuetan hobekuntzak lortu ziren, beste kasu batzuetan hobekuntza horiek norabide bakarrean eman ziren (unitate mota aldatzean), eta beste batzuetan proposatutako aldaketak ez zuen hobekuntzarik ekarri. Edonola ere, esperimentu guztiak kontuan hartuta, aurretik optimizatutako sistema bat (Etchegoyhen *et al.*, 2018) are gehiago optimizatzea lortu zen, eu-es zentzuan proba multzoan lortutako emaitzak 0,47 BLEU hobetuz eta es-eu zentzuan 0,86 BLEU irabaziz.

Sistema hauek ondoren domeinu klinikora egokitzeko egindako proben oinarri bezala erabili ziren, eta beraz domeinuz kanpoko ebaluazio multzo hauek ez ziren berriro erabili (horrek justifikatzen du erabakiak hartzeko

Noranzkoa	**Hiperparametroaren balioa**	**dev BLEU**	**test BLEU**
eu-es	Optimizatzailea = Adadelta	26,51	28,98
	Optimizatzailea → Adam	**26,87**	**28,97**
	Unitate mota = GRU	**26,87**	**28,97**
	Unitate mota → LSTM	26,87	28,68
	Beam-width = 6	26,87	28,97
	***Beam-width* → 10**	**27,21**	**29,28**
	Esaldi multzoaren tamaina = 30	27,21	29,28
	Esaldi multzoaren tamaina → 32	27,08	29,48
	Esaldi multzoaren tamaina → 64	**27,02**	**29,45**
	Hitz-bektorearen tamaina = 500	**27,02**	**29,45**
	Hitz-bektorearen tamaina → 512	26,65	28,87
es-eu	Optimizatzailea = Adadelta	22,95	20,26
	Optimizatzailea → Adam	**23,06**	**20,55**
	Unitate mota = GRU	23,06	20,55
	Unitate mota → LSTM	**23,37**	**20,96**
	Beam-width = 6	23,37	20,96
	***Beam-width* → 10**	**23,64**	**20,93**
	Esaldi multzoaren tamaina = 30	23,64	20,93
	Esaldi multzoaren tamaina → 32	22,88	20,65
	Esaldi multzoaren tamaina → 64	**23,05**	**21,12**
	Hitz-bektorearen tamaina = 500	**23,05**	**21,12**
	Hitz-bektorearen tamaina → 512	23,09	20,42
	Hitz-bektorearen tamaina → 1024	23,09	20,61

5.2 taula – eu-es eta es-eu norabideetan probatutako hiperparametroen balio desberdinekin domeinuz kanpoko garapen eta proba multzoetan lortutako BLEU balioak. Hiperparametroaren balioa oinarri-lerro sistemarekiko mantentzen denean '=' sinboloa erabiltzen da, eta aldatzen denean '→'.

proba multzoan eta ez soilik garapen multzoan lortutako emaitzak kontuan hartzea).

5.2 Domeinu klinikora egokitzeko lehenengo saiakerak

Helburua

Domeinuz kanpo optimizatutako sistemak domeinu klinikora egokitzea, domeinuko baliabide desberdinen eragina neurtuz.

Corpusak

Jakina denez, IAN sistema batek emaitza kaxkarrak lortuko ditu entrenamendu-corpusean agertzen ez den domeinu batetako esaldiak itzultzerakoan. Hortaz, domeinu klinikoko itzultzaile automatiko bat garatu nahi badugu, beharrezkoa izango da entrenamendu-corpusera domeinu klinikoko testuak txertatzea. Helburu horrekin egindako lehenengo esperimentu multzoan, aurreko atalean domeinuz kanpo optimizatutako sistema erabili zen oinarri moduan, eta sistema hori entrenatzeko erabilitako corpusera domeinu klinikoko testuak gehitzen joan zitzaizkion.

Zehazki, bi baliabide desberdin gehitu ziren domeinu klinikora egokitzeko lehenengo esperimentu hauetan: SNOMED CTren euskaratzeari esker lortutako terminologia klinikoaren lehenengo bertsioa (gaztelerazko termino baliokideekin batera), eta Galdakao-Usansolo Ospitaleko gaztelerazko alta-txostenak. Baliabide bakoitza bi modu desberdinetan gehitu zen: terminologia klinikoaren kasuan, modu zuzenean eta SNOMED CTn definitutako erlazioetan oinarrituta sortutako esaldi artifizialen bidez; eta gaztelerazko alta-txostenen kasuan, atzeranzko itzulpena (Sennrich *et al.*, 2016) eta kopiatze (Currey *et al.*, 2017) tekniken bidez. Esperimentu hauen helburua domeinu klinikora egokitzeko gaitasuna neurtzea izanda, ebaluaziorako garaian eskuragarri zeuden euskarazko alta-txosten ereduak (Joanes Etxeberri Saria V. Edizioa, 2014) eta haien eskuzko itzulpenak erabili ziren. Laburpen moduan, 5.3. taulak atal honetan deskribatutako sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak aurkezten ditu.

Esperimentuak

Atal honetan aurkeztutako esperimentuetarako 3 sistema garatu ziren es-eu zentzuan, eta 5 sistema eu-es zentzuan, aipatutako baliabideak bata bestea-

Corpusa	**Esaldiak**
Domeinuz kanpoko 1-7 corpusak	4.530.683
SNOMED CT 1.0 terminoak	151.111
SNOMED CTtik eratorritako esaldi artifizialak	363.958
Galdakao-Usansolo Ospitaleko gaztelerazko corpusa	2.023.811

5.3 taula – Domeinu klinikora egokitzeko lehenengo saiakeretan sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak

ren ondoren gehituz. Bi norabideetan, oinarri-lerro moduan aurreko atalean domeinuz kanpoko 1-7 corpusekin entrenatutako sistemak domeinu klinikoan ebaluatu ziren. Ondoren, norabide bakoitzean SNOMED CT 1.0 terminologia gehituz beste bi sistema garatu ziren; eta jarraian, terminologia beretik sortutako esaldi artifizialak gehituta sistema bana entrenatu zen. Gainera, eu-es zentzuan beste sistema bat entrenatu zen Galdakao-Usansolo Ospitaleko gaztelerazko alta-txostenak atzeranzko itzulpen bidez gehituta; eta azkeneko sistema bat corpus bera kopiatze teknikaren bidez ere gehituta. Sistema guztiak entrenatzeko *Nematus* erabili zen, aurreko atalean optimizatutako hiperparametroekin. Atzeranzko itzulpena egiteko, es-eu zentzuan SNOMED CT 1.0 terminologia modu zuzenean gehitu ondoren entrenatutako sistema erabili zen. Dekodetze-teknika bezala, *beam-search* erabili zen, aurreko ataleko emaitzak kontuan hartuta *beam-width*-a 10 izanda.

Jarraian, esaldi artifizialak sortzeko jarraitutako metodologia azalduko da. Erlazioak erauzteko, 2017ko uztailaren 31n argitaratutako SNOMED CTren *Snapshot* bertsioa erabili zen, RF2 formatuan. SNOMED CT kontzeptuka antolatuta dago eta kontzeptu bakoitza zein terminoekin deskribatzen den azaltzen da. Horretaz gain, kontzeptuak erlazioen bitartez lotzen dira, eta guraso/ume erlazioez gain bestelako erlazioak ere badaude; adibidez, '*causative agent*' edo '*procedure site*'. Sortutako esaldiak esanguratsuak izan zitezen, SNOMED CTn definitutako erlazio aktibo guztietatik ohikoenak bakarrik erabili ziren, zehazki 10.000 aldiz baino gehiagotan agertzen ziren erlazioak kontuan hartuz.

Behin erlazio motak zehaztuta, bakoitzari zegozkien bi esaldi mota definitu ziren euskaraz eta gazteleraz. Modu honetan, 5.4. eta 5.5. taulek erlazio mota ohikoenak eta bakoitzarentzako sortutako esaldi artifizialen ereduak erakusten ditu, erlazioa osatzen duten terminoak adierazteko X eta Y aldagaiak erabiliz. Esaldi hauek eredu bezala hartuta, esperimentu hauetan erabilitako SNOMED CTren lehenengo bertsioan agertzen den X termino ele-

bidun bakoitzarentzat, termino honek dituen erlazio aktibo guztietatik bat ausaz aukeratu zen, erlazio ohikoenetara mugatuz eta erlazioa osatzen duen Y terminoa itzulita dagoela baldintza bezala hartuz. Azkenik, erlazio horretarako definitutako 2 esaldi-ereduetatik bat ausaz aukeratu zen, eta erlazio horri zegozkien euskarazko eta gaztelerazko esaldiak gehitu ziren sortutako corpusera.

Behin esaldi artifizial hauek sortuta, erlazio mota bakoitzerako sortutako ereduak berrikusi ziren, eta termino gehienetarako sortutako esaldiak egokiak ez baziren erlazio mota horri zegozkien esaldi guztiak baztertu ziren. Modu horretan, *Occurrence* eta *Direct substance* erlazioei zegozkien esaldiak baztertzea erabaki zen, esaldi-ereduak berridazteko modu egokirik topatzea ezinezkotzat hartu ondoren. Gainera, termino elebidun bakoitzarentzat erlazio bat ausaz aukeratzerakoan, *Subject relationship context* erlaziorako ez zen esaldi bat bera ere sortu.

Azkenik, 5.4. eta 5.5. tauletan agertzen diren inflexio morfologikoak modu egokian aplikatu ahal izateko Xuxen (Agirre *et al.*, 1992) zuzentzailean erabiltzen diren inflexio-erregelak aplikatu ziren. Modu honetan, domeinuz kanpoko 1-7 corpusei eta SNOMED CTren euskaratzetik sortutako lehenengo bertsioari 363.958 esaldi gehiago batu zitzaizkien. Esan beharra dago, SNOMED CTko termino hutsak gehitzeaz gain, esaldi artifizial hauek gehitzeko helburua IAN sistemen ikasketa-prozesuan laguntzea izan zela, dekodetze-prozesuan hurrengo hitza sortzeko erabiltzen den hizkuntza-eredu modukoari terminoak testuinguru batean emanez.

Emaitzak

Domeinu klinikora egokitzeko egindako lehenengo esperimentu multzoaren emaitzak 5.6. taulan aurkezten dira. Hiperparametroen optimizazioarekin egin moduan, ebaluaziorako BLEU metrika erabili zen, eta emaitzak garapen eta proba multzoetan erakusten dira, kasu honetan ebaluaziorako esaldi hauek Donostia Unibertsitate Ospitaleko alta-txosten ereduak eta haien gaztelerazko eskuzko itzulpenak izanik. Proba (*test*) corpusean emaitza hoberenak letra lodiz azaltzen dira.

Erlazioa	**Euskarazko esaldi-eredua (I)**	**Gaztelerazko esaldi-eredua (I)**
is a	X Y da	X es Y
Finding site	X Yn gertatzen da	X ocurre en Y
Associated morphology	X Y moduan gertatzen den gaixotasuna da	X es una enfermedad que ocurre en forma de Y
Method	X Y behar duen prozedura da	X es un procedimiento que requiere de Y
Procedure site - Direct	X Y zuzenean ukitzen duen prozedura da	X es un procedimiento que afecta directamente a Y
Procedure site	X Y ukitzen duen prozedura da	X es un procedimiento que afecta a Y
Part of	X Yren parte bat da	X es una parte de Y
Interprets	X Y interpretatzen duen aurkikuntza da	X es un hallazgo que interpreta Y
Causative agent	X Yk eragindako gaixotasuna da	X es una enfermedad causada por Y
Direct morphology	X metodoa aplikatzean Y zuzenean ukitzen da	Al aplicar el método X se afecta directamente a Y
Procedure site - Indirect	X Y zeharka ukitzen duen prozedura da	X es un procedimiento que afecta indirectamente a Y
Has active ingredient	X produktuak Y substantzia dauka	El producto X tiene la sustancia Y
Has interpretation	X aurkikuntza Y bezala interpretatzen da	El hallazgo X se interpreta como Y
Temporal context	X Y kokatzen da	X se sitúa Y
Subject relationship context	X egoerak Y pertsonari eragiten dio	La situación X afecta a la persona Y
Occurrence	X lehenengoz Yn azaldu zen	X apareció por primera vez en Y
Direct substance	X prozedurak Y substantzia erabiltzen du	El procedimiento X utiliza la sustancia Y
Pathological process	X gaixotasunak Y suposatzen du	La enfermedad X supone Y
Has manufactured dose form	X produktua Y bezala banatzen da	El producto X se reparte como Y

5.4 taula – Esaldi artifizialak sortzeko erabilitako esaldi-ereduak (I)

Ondorioak

Ikus dezakegunez, gehitutako baliabide bakoitzak BLEU balioan izandako eragina oso desberdina izan zen. Ekarpen handiena ezaguna zen atzeranzko itzulpenak ekarri zuen, proba multzoan 5,59 BLEU puntu igoz. Honen on-

Erlazioa	**Euskarazko esaldi-eredua (II)**	**Gaztelerazko esaldi-eredua (II)**
is a	X Y mota bat da	X es un tipo de Y
Finding site	X Yn aurkitzen da	X se encuentra en Y
Associated morphology	X Y forma hartzen duen gaixotasuna da	X es una enfermedad que toma la forma de Y
Method	X prozedurak Y suposatzen du	El procedimiento X supone Y
Procedure site - Direct	X prozedura praktikatzean Y zuzenean ukitzen da	Al practicar el procedimiento X se afecta directamente a Y
Procedure site	X prozedura praktikatzean Y ukitzen da	Al practicar el procedimiento X se afecta a Y
Part of	X Yren parte da	X forma parte de Y
Interprets	X aurkikuntzak Y interpretatzen du	El hallazgo X interpreta Y
Causative agent	X Yren eraginez sortutako gaixotasuna da	X es una enfermedad causada por efecto de Y
Direct morphology	X metodoak Y zuzenean ukitzen du	El método X afecta directamente a Y
Procedure site - Indirect	X prozedura praktikatzean Y zeharka ukitzen da	Al practicar el procedimiento X se afecta indirectamente a Y
Has active ingredient	X Y substantzia daukan produktua da	X es un producto que contiene la sustancia Y
Has interpretation	X aurkikuntzak Y interpretazioa du	El hallazgo X tiene la interpretación Y
Temporal context	X denboran Y kokatzen da	X se sitúa temporalmente Y
Subject relationship context	X egoera Y pertsonari dagokio	La situación X le corresponde a la persona Y
Occurrence	X Yn agertu zen	X surgió en Y
Direct substance	Y substantzia X prozeduran erabiltzen da	La sustancia Y se utiliza en el procedimiento X
Pathological process	X Y moduan agertzen den gaixotasuna da	X es una enfermedad que aparece en forma de Y
Has manufactured dose form	X produktua Y moduan banatzen da	El producto X se reparte en forma de Y

5.5 taula – Esaldi artifizialak sortzeko erabilitako esaldi-ereduak (II)

doren, nabarmena da SNOMED CTko terminologia klinikoa gehitzeak egindako ekarpena, eu-es zentzuan proba multzoan 4,37 puntuko igoerarekin.

Noranzkoa	Entrenamendu-corpusa	dev BLEU	test BLEU
eu-es	Domeinuz kanpokoa (1-7)	10,69	10,67
	+ SNOMED CT 1.0	15,45	15,04
	+ esaldi artifizialak	16,08	15,48
	+ atzeranzko itzulpena	22,52	21,07
	+ kopiatzea	23,57	**21,59**
es-eu	Domeinuz kanpokoa (1-7)	9,08	8,69
	+ SNOMED CT 1.0	10,75	**10,44**
	+ esaldi artifizialak	10,79	10,43

5.6 taula – eu-es eta es-eu norabideetan domeinu klinikoko baliabideak gehitzean Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak

Igoera hau txikiagoa da es-eu zentzuan, ziurrenik gaztelerazko terminoak eskuz itzuliak eta euskarazkoak automatikoki itzulitakoak izateagatik. Edonola ere, SNOMED CTko terminoak modu zuzenean entrenamendu-corpusera gehitzean lortutako emaitzek erakusten dute, sarreran adierazten genuen hipotesiari jarraituz, osasun arloko itzulpenean terminologiak garrantzi handia duela. Honek tesi honen lehenengo ondorio azpimarragarria uzten digu:

1. Domeinuko testu elebidunik gabeko egoera batean, terminologia klinikoa modu zuzenean entrenamendu-corpusera gehitzea lagungarria da.

Bestalde, esaldi artifizialak gehitzeak uste baino eragin txikiagoa izan zuen, eu-es zentzuan igoera txikia sortuz (+0,44 proba multzoan) eta es-eu zentzuan garapen eta proba multzoetan lortutako emaitzak modu esanguratsuan aldatu gabe. Hau kontuan hartuta, corpusa eguneratzean egindako esperimentuetarako teknika hau erabiltzea baztertu egin zen, esaldi artifizialak sortzeko definitutako prozesuak duen berezko ausazkotasuna ere kontuan izanik. Gainera, kontuan hartu behar da termino desberdinak txertatzeko esaldi-eredu berdinak erabiltzeak sistemak nahastu ditzakeela, termino desberdin hauek testuinguru berdinean agertzean ikasitako hitz-bektoreak antzekoak izango direlako.

Azkenik, corpus erlatiboki txikia erabili dugun egoera honetan, kopiatze teknika ere lagungarria izan daitekeela erakusten da, nahiz eta aurretik

corpus bera atzeranzko itzulpenaren bidez gehitu izan. Gure hipotesia da teknika honek batez ere euskaraz eta gazteleraz berdin idazten diren entitate izendunak itzultzeko lagungarria dela (adib.: medikamentuen izenak).

Giza ebaluazioa

Lehenengo esperimentu multzo hau amaitzeko, eu-es zentzuan domeinu klinikoko baliabide guztiak gehituta BLEU altuena lortu zuen sistemaren giza ebaluazioaren emaitzak aurkezten dira, ebaluazio hau Osakidetzako bi mediku elebidunek egin zutelarik. Horretarako, proba multzotik ausaz erauzitako 100 esaldi erabili ziren lagin moduan, eta hauen gainean bi motatako ebaluazioak egin ziren: 1) gure sistemaren irteera eskuzko itzulpenarekin eta garaiko *Google Translate*[1] sistemaren itzulpenarekin alderatzea, itzulpenen arteko sailkapena eginez; eta 2) gure sistemen itzulpenen zehaztasuna eta naturaltasuna neurtzea.[2]

Itzulpen desberdinen sailkapena irudikatzeko, bi alderaketa egin genituen: 1) eskuzko itzulpenen eta gure sistemaren itzulpenen artean; eta 2) gure sistemaren eta *Google Translate*-en itzulpenen artean. Lehenik eta behin, 5.7. taulak proba multzotik hartutako 100 esaldien sailkapenen distribuzioa erakusten du, baldin eta eskuzko itzulpena (esk.) gure sistemaren (aut.) itzulpena baino hobea, berdina edo okerragoa bezala sailkatu zen. Ondoren, 5.8. taulak emaitza berdinak erakusten ditu gure sistemaren eta *Google Translate*-en itzulpenen arteko alderaketa irudikatzeko. Anotatzaileen arteko adostasuna neurtzeko *Cohen-en kappa* neurtu zen, eta honek 0,25 balioa hartu zuen eskuzko itzulpena eta gure sistemaren arteko alderaketan (adostasun nahikoa), eta 0,17 balioa gure sistemaren eta *Google Translate*-en arteko alderaketan (adostasun arina).

Bi taula hauetako emaitzak kontuan hartuta, argi ikusten da gure sistema *Google Translate* baino hobea zela osasun-txostenak euskaratik gaztelerara itzultzeko; eta gure sistemaren itzulpena eta eskuzko itzulpenaren artean, eskuzkoa hobetzat hartzen zen, nahiz eta emaitzak parekatuagoak izan. Edonola ere, giza ebaluazio honen emaitzak baloratzerakoan kontuan hartu behar da, 4.1.3. azpiatalean aipatu den moduan, ebaluazio-corpus honen ezaugarriak bereziak direla, eta jatorrizko euskarazko alta-txosten ereduak

[1] `https://translate.google.com/`

[2] Errore mota desberdinak ere aztertu ziren, baina hemen laburpen moduan aurrekoak bakarrik aurkeztuko dira.

Ebaluatzailea	**esk. > aut.**	**esk. = aut.**	**esk. < aut.**
1. ebaluatzailea	42	31	27
2. ebaluatzailea	38	30	32
Batez bestekoa	40	30,5	29,5

5.7 taula – Proba multzoko 100 esaldien sailkapenen distribuzioa, baldin eta eskuzko itzulpena (esk.) gure sistemaren (aut.) itzulpena baino hobea, berdina edo okerragoa bezala sailkatu zen

Ebaluatzailea	**aut. > Google**	**aut. = Google**	**aut. < Google**
1. ebaluatzailea	64	18	18
2. ebaluatzailea	49	31	20
Batez bestekoa	56,5	24,5	19

5.8 taula – Proba multzoko 100 esaldien sailkapenen distribuzioa, baldin eta gure sistemaren itzulpena (aut.) *Google Translate*-en (Google) itzulpena baino hobea, berdina edo okerragoa bezala sailkatu zen.

ondo osatuak dauden moduan, mediku batek egindako gaztelerazko itzulpenek askotan esaldiak laburtu egiten dituztela eta akats ortografikoak izan ohi dituztela. Hortaz, proba multzoko esaldi batean euskarazko esaldian gaztelerazkoan agertzen ez den terminoren bat agertzen bada, edota gaztelerazko itzulpen honek akats ortografikoren bat badu, errazagoa izango da gure sistemak sortutako itzulpena eskuzko itzulpena baino hobea bezala sailkatzea.

Gure sistemen kalitatea hobeto ebaluatzeko, ebaluatzaile berdinei itzulpenen zehaztasuna eta naturaltasuna 1etik 4rako eskalan kokatzeko eskatu zitzaien, 1 baliorik txarrena eta 4 baliorik hoberena izanik. Honela, 5.9. eta 5.10. taulek proba multzotik hartutako 100 esaldien zehaztasuna eta naturaltasunari emandako balioen distribuzioa erakusten dute, hurrenez hurren.

Orokorrean, itzulpen gehienek balorazio ona jaso zutela ikus daiteke; eta IAN sistemetan ohikoa den moduan, naturaltasunari emandako balorazioak zehaztasunari emandakoak baino altuagoak dira gure sisteman ere. Anotatzaileen arteko adostasunari dagokionez, zehaztasunaren kasuan *Cohen-en kappa*-k 0,15 balioa hartu zuen (adostasun arina); eta naturaltasunaren kasuan, 0,65 (adostasun handia).

Ebaluatzailea	**4**	**3**	**2**	**1**
1. ebaluatzailea	50	36	12	2
2. ebaluatzailea	57	29	12	2
Batez bestekoa	53,5	32,5	12	2

5.9 taula – Proba multzoko 100 esaldien zehaztasunari emandako balioen distribuzioa (4: hoberena; 1: txarrena)

Ebaluatzailea	**4**	**3**	**2**	**1**
1. ebaluatzailea	88	11	0	1
2. ebaluatzailea	80	13	6	1
Batez bestekoa	84	12	3	1

5.10 taula – Proba multzoko 100 esaldien naturaltasunari emandako balioen distribuzioa (4: hoberena; 1: txarrena)

5.3 Sistema desberdinak alderatu eta atzeranzko itzulpena egiteko teknika desberdinak probatzea

Transformer (Vaswani *et al.*, 2017) arkitekturaren agerpenarekin, beharrezkoa izan zen honekin lortutako emaitzak aurretik garatutako SNEarekin lortutakoekin alderatzea. Behin alderaketa hau eginda, geruza gehiago zituen SNE bat (Barone *et al.*, 2017) ere probatu zen. Honekin batera, atzeranzko itzulpena egiteko aurretik garatutako SNEa eta Transformer arkitekturaz gain, EOIA eta IAE sistemak ere probatu ziren, baliabide gutxiagoko egoera batean baliagarriak izan zitezkeen aztertzeko.

Helburua

IAN sistema desberdinak alderatzea, eta atzeranzko itzulpena egiteko EOIA eta IAE ere baliagarriak izan daitezkeen aztertzea.

Corpusak

Esperimentu hauek egiteko aurretik erabilitako corpus berdina erabili zen: domeinuz kanpoko 1-7 corpusak, SNOMED CTren euskaratzetik erauzitako

lehenengo bertsioa; eta eu-es zentzuan, SNOMED CTren erlazioetatik sortutako esaldi artifizialak, Galdakao-Usansolo Ospitaleko alta-txostenak atzeranzko itzulpenaren bidez eta kopiatze bidez ere gehituak. Corpus hauen esaldi kopuruak 5.3. taulan ikus daitezke.

Esperimentuak

Egin beharreko esperimentuak bi pausotan aurrera eraman ziren: 1) eu-es zentzuan, atzeranzko itzulpena egiteko aurretik garatutako SNEa erabilita, SNE sakona (Barone *et al.*, 2017) eta Transformer (Vaswani *et al.*, 2017) arkitekturekin lortutako emaitzak aurretik garatutako SNEarekin lortutakoekin alderatu; eta 2) lehenengo esperimentuan emaitza automatiko hoberenak lortzen zituen arkitekturarekin, atzeranzko itzulpena egiteko Transformer arkitektura zein EOIA eta IAE sistemak ere probatu.

eu-es zentzuan egin beharreko esperimentuetarako, Transformer arkitekturaren *OpenNMT*-ren (Klein *et al.*, 2017) inplementazioa erabili zen, garaian gomendatutako hiperparametroekin; eta SNE sakonaren kasuan (Barone *et al.*, 2017), bertan probatutako sistemen artean emaitza hoberenak ematen zituen konfigurazioa erabili zen.[3]

Atzeranzko itzulpena egiteko sistema desberdinak probatu ziren. EOIAren kasuan *MatxinMed* sistema erabili zen, IAE egiteko *Moses* (Koehn *et al.*, 2007) tresna erabili zen, 4.2.2. azpiatalean adierazitako konfigurazioarekin, eta Transformer arkitekturarako kontrako zentzuan erabilitako *OpenNMT*-ren (Klein *et al.*, 2017) inplementazio berdina erabili zen. IAE eta Transformer sistemak entrenatzeko aurretik garatutako SNEak erabilitako corpus berdina erabili zen, domeinuz kanpoko 1-7 corpusek eta SNOMED CT 1.0 bertsioak osatua. Transformerraren kasuan, dekodetze-teknika bezala *unrestricted sampling* (Edunov *et al.*, 2018) erabili zen, garaian artearen egoera bilakatu zena.

Emaitzak

Honela, 5.11. taulak eu-es zentzuan arkitektura desberdinekin Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak erakusten ditu, atzeranzko itzulpena egiteko

[3] `https://github.com/Avmb/deep-nmt-architectures/blob/master/configs/bideep-bideep-rGRU-large/config.sh`

aurretik garatutako SNEa erabilita. Emaitza hoberenak letra lodiz agertzen dira.

Arkitektura	**dev BLEU**	**test BLEU**
SNE	23,57	21,59
SNE sakona	23,01	20,74
Transformer	**26,67**	**24,44**

5.11 taula – eu-es zentzuan arkitektura desberdinekin Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak, atzeranzko itzulpena egiteko aurretik garatutako SNEa erabilita.

Jarraian, 5.12. taulan es-eu zentzuan atzeranzko itzulpena egiteko sistema bakoitzarekin Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak erakusten dira. Emaitza hoberenak letra lodiz agertzen dira.

Sistema	**dev BLEU**	**test BLEU**
EOIA	8,56	7,03
IAE	10,30	8,75
SNE	10,75	10,44
Transformer	**11,30**	**12,04**

5.12 taula – es-eu zentzuan sistema desberdinekin Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak

Azkenik, 5.13. taulak eu-es zentzuan Transformer arkitekturarekin atzeranzko itzulpena egiteko sistema desberdinak erabilita Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak erakusten ditu. Emaitza hoberenak letra lodiz agertzen dira.

Ondorioak

eu-es zentzuan, atzeranzko itzulpena egiteko aurretik garatutako SNEa erabilita, 5.11. taulan ikus daitekeen moduan, Transformer arkitekturak nabarmen hobetzen ditu SNEarekin lortutako emaitzak, proba multzoan BLEU

Sistema	**dev BLEU**	**test BLEU**
EOIA	22,98	21,91
IAE	22,78	21,43
SNE	26,67	24,44
Transformer	**27,70**	**25,61**

5.13 taula – eu-es zentzuan Transformer arkitekturarekin atzeranzko itzulpena egiteko sistema desberdinak erabilita Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako garapen eta proba multzoetan lortutako BLEU balioak

balioa 2,85 puntu igoaraziz. Aurreikusi litekeenaren kontrara, 4 geruzako SNE sakonarekin aurretik garatutako geruza bakarreko SNEarekin lortutako emaitzak ez ziren hobetu. Hau azaltzeko arrazoiak bi izan daitezke: 1) entrenamendu-corpuseko esaldi gehienak domeinuz kanpokoak izanda, SNE sakonak *overfitting* edo gaindoitze arazoak izatea, osasun arloko ebaluazio-corpusean emaitza okerragoak lortuz; eta 2) aurretik garatutako SNEaren hiperparametroak erabilitako corpuserako optimizatu izana.

Atzeranzko itzulpena egiteko sistemei dagokienez, 5.12. taulan ikusten denez, es-eu zentzuan ere Transformer arkitekturak lortzen ditu emaitza hoberenak, nahiz eta kasu honetan IAE eta SNE sistemek antzeko emaitzak lortu, batez ere garapen multzoan. Edonola ere, guri gehien interesatzen zaiguna ez da sistema desberdinek es-eu zentzuan lortzen dituzten emaitzak zeintzuk diren ezagutzea, baizik eta atzeranzko itzulpena egiteko sistema hauek erabilita eu-es zentzuan Transformer arkitektura erabiliz lortzen diren emaitzak ezagutzea.

Modu honetan, 5.13. taulan ikusten denez, atzeranzko itzulpena egiteko ere Transformer arkitektura da eraginkorrena, aurreko SNEarekin alderatuta 1,17 BLEU irabaziz. Emaitza hoberenak zein diren alde batera utzita, esanguratsua da EOIA sistemarekin IAEren emaitzak hobetu izana, nahiz eta, 5.12. taulan ikusten denez, es-eu zentzuan IAE sistemak emaitza hobeak lortu. Honen azalpen moduan, kontuan hartu behar da BLEU metrikak EOIA sistemen kalitatea gutxiesten duela (Mayor, 2007).

Atal hau amaitzeko, esperimentu multzo honetatik ateratako bi ondorio nabarmentzen dira:

2. Osasun-txostenak euskararen eta gazteleraren artean itzultzeko Transformer da arkitektura hoberena, baita atzeranzko itzulpena egiteko ere.

3. Domeinuko corpus elebidunik gabeko egoera batean, EOIA aproposa izan daiteke atzeranzko itzulpena egiteko, IAEren emaitzak hobetuz.

5.4 Atzeranzko itzulpenerako sistemak alderatu, konbinatu, sortutako corpus sintetikoen aniztasun lexikala aztertu eta datuen hautespena aplikatzea

Atal honetan, euskararen eta gazteleraren artean txosten klinikoak itzultzeko sistema ahalik eta hoberena garatzeko helburutik haratago joan ginen, atzeranzko itzulpena, aniztasun lexikala eta datuen hautespena uztartzeko asmoz.

Helburua

Atzeranzko itzulpena egiteko sistema desberdinen irteera konbinatzea lagungarria izan daitekeen aztertzea, sortutako corpus sintetikoen aniztasun lexikala neurtzea eta hauei datuen hautespena aplikatzearen eragina ikertzea.

Corpusak

Esperimentu hauetarako, aurretik erabilitako euskara / gaztelera corpusez gain, antzeko ezaugarriak zituzten alemana / ingelesa (aurrerantzean, de/en) corpusak erabili ziren. Euskara eta gaztelerazko corpusen kasuan, aurretik erabilitako corpus berak erabili ziren, esaldi artifizialak baztertuta eta ebaluaziorako esaldietatik errepikatuak zeudenak ezabatuta. Modu honetan, aurretik Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako 2.076 esaldi pare elebidunetatik 1.648 esaldi ez-errepikatu geratu ziren, horietatik 824 garapenerako eta beste 824 probarako erabiliz. Sistemak entrenatzeko

erabili ziren corpusen esaldi kopuruak (baztertutako esaldi artifizialak barne) 5.3. taulan ikus daitezke.

Alemana eta ingelesaren artean egin beharreko esperimentuetarako, domeinuz kanpoko corpus bezala WMT 2015 (Bojar *et al.*, 2015) konferentzian albisteak itzultzeko atazan erabilitako 4,5M esaldiko corpusa erabili zen, eta atzeranzko itzulpena egiteko domeinuko corpus moduan biomedikuntzaren arloko UFAL[4] corpusaren ingelesezko zatia erabili zen, 2,3M esaldiekin. Ebaluaziorako, domeinu klinikoko HimL[5] corpusa erabili zen; zehazki Cochrane corpuseko 467 esaldiak garapenerako eta NHS corpuseko 1.044 esaldiak probarako erabiliz.

Esperimentuak

Egin beharreko esperimentuak lau urratsetan banatu ziren:

1. Corpus elebidunarekin eu-es eta de-en zentzuetarako LSTM eta Transformer sistemak entrenatu *OpenNMT* (Klein *et al.*, 2017) erabiliz, 4.2. atalean aipatutako hiperparametroekin.

2. Atzeranzko itzulpena egiteko es-eu eta en-de zentzuetan 4.2. atalean aipatutako EOIA, IAE, LSTM eta Transformer sistemak erabili, beraien IA metrikak kalkulatuz eta sortutako corpusetan aniztasun lexikala neurtuz. LSTM eta Transformer bidez itzulpenak sortzeko, *beam-search* teknika erabili zen dekodetze-teknika moduan, *beam-width* 5 izanda.

3. Lehenengo urratsean emaitza hoberenak lortzen dituen sistemarekin, eu-es eta de-en zentzuetan sistema bana entrenatu bigarren pausoan atzeranzko itzulpenaren bidez sortutako corpus bakoitzarekin eta guztiekin batera, lortutako IA metrikak alderatuz.

4. Atzeranzko itzulpenaren bidez sortutako 4 corpusek batera osatutako multzoan datuen hautespena aplikatu, corpus bakoitzaren tamaina bereko corpusa hautatuz, eta modu honetara erauzitako corpusa corpus elebidunera gehitu, aurreko pausoaren antzera eu-es eta de-en zentzuetan sistema bana entrenatuz.

[4] `https://ufal.mff.cuni.cz/ufal_medical_corpus`
[5] `https://www.himl.eu/test-sets`

IA sistemak ebaluatzeko, 2. kapituluan aipatutako lau metrikak kalkulatu ziren (BLEU, TER, METEOR eta chrF3), eta atzeranzko itzulpenaren bidez sortutako corpusen aniztasun lexikala aztertzeko, kapitulu berean deskribatutako hiru metrikak neurtu ziren (TTR, *Yules' I* eta MTLD).

Era berean, datuen hautespena aplikatzeko 4 modu desberdin probatu ziren:

1. *FromAll*: Atzeranzko itzulpenaren bidez sortutako 4 corpusen multzotik, datuen hautespenean balio altuenak lortzen dituzten esaldiak aukeratu, esaldi bera sistema desberdinek itzulita hautatzeko aukera emanez.

2. *EachFromAll*: Atzeranzko itzulpena aplikatu zaion corpusean agertzen den esaldi bakoitzerako, sistema desberdinak erabiliz sortutako 4 esaldietatik 1 aukeratu.[6]

3. *EachFromAll x4*: Aurrekoaren berdina, baina hautatutako corpusa 4 aldiz errepikatuta, atzeranzko itzulpena egiteko sistema guztiek sortutako corpusarekin egindako esperimentuarekin alderatzeko.[7]

4. *EachFromAll RS*: Datuen hautespena aplikatzerakoan esaldi bakoitzarentzako lortutako balioa birkalkulatu, esaldi hori itzultzeko erabili den sistemari dagozkion IA eta aniztasun lexikala neurtzeko metrikak faktore moduan erabiliz, eta berrordenatutako zerrendan *EachFromAll* moduko hautapena egin, atzeranzko itzulpena aplikatu zaion corpusean agertzen den esaldi bakoitzerako itzulpen bakarra hautatuz.

Azken hau aurrera eramateko, hainbat proba egin ziren IA eta aniztasun lexikala neurtzeko metrika desberdinekin, modu enpirikoan es-eu eta en-de zentzuetan IA emaitza hoberenak lortzen zituzten sistemek itzulitako esaldi gehien hautatzen zituen konbinazioa erabiliz. Modu horretan, datuen hautespenaren bidez lortutako emaitzak birkalkulatzeko faktore moduan honako balioa erabili zen, datuen hautespenean esaldi bakoitzari emandako balioari biderkatuz aplikatu zena, esaldi hori itzultzeko erabili zen sistemaren metriken arabera: $\phi = log(BLEU * (100 - TER) * MTLD)$

[6]Kontuan hartuta erabilitako FDA algoritmoak ez diela baliorik ematen garapen multzoko esaldiekiko hitz-segida komunik ez dituzten esaldiei, 4 sistemek sortutako itzulpenetan hau gertatzen zenean ausaz aukeratzen zen 4 esaldietatik 1.

[7]Esperimentu hau eu-es zentzuan bakarrik egin zen.

Sistema desberdinen emaitzak erakutsi baino lehen, aipatu beharra dago datuen hautespena aplikatu ondoren entrenatutako IAN sistemak *early stopping* teknikaren bidez entrenatu zirela, *perplexity* balioa 3 pauso jarraituetan txikitzen ez zenean entrenamendua geldituz eta gordetako azken sistemarekin proba multzoan IA metrikak neurtuz. Ostera, datuen hautespena aplikatu baino lehen entrenatutako sistemak, tesi honetan *OpenNMT*-rekin entrenatutako beste sistema guztiak bezala, 200.000 pausoz entrenatu ziren, garapen multzoan BLEU altuena lortzen zuen sistema hautatuz berarekin proba multzoan IA metrikak neurtzeko.

Emaitzak

Hasteko, 5.14. taulak eu-es eta de-en zentzuetan corpus elebidunekin entrenatutako sistemen IA emaitzak erakusten ditu.

	Sistema	**BLEU↑**	**TER↓**	**METEOR↑**	**chrF3↑**
EU-ES	LSTM	10,84	85,00	32,79	41,36
	Transformer	**19,64**	**69,11**	**43,84**	**53,03**
DE-EN	LSTM	28,15	51,95	32,19	55,40
	Transformer	**38,27**	**42,87**	**37,02**	**62,37**

5.14 taula – eu-es eta de-en zentzuetan corpus elebidunekin entrenatutako sistemen IA emaitzak

Ondoren, 5.15. taulan es-eu eta en-de zentzuetan atzeranzko itzulpena egiteko sistema desberdinek lortutako IA emaitzak aurkezten dira.

	Sistema	**BLEU↑**	**TER↓**	**METEOR↑**	**chrF3↑**
ES-EU	EOIA	11,37	75,52	19,80	41,35
	IAE	9,38	70,70	25,36	44,07
	LSTM	7,01	72,29	20,46	33,94
	Transformer	**12,21**	**66,53**	**26,96**	**44,42**
EN-DE	EOIA	8,21	72,26	25,70	41,40
	IAE	14,85	74,00	35,62	48,92
	LSTM	24,65	54,60	43,30	53,51
	Transformer	**32,24**	**46,83**	**50,25**	**60,29**

5.15 taula – es-eu eta de-en zentzuetan atzeranzko itzulpena egiteko sistemek lortutako IA emaitzak

Jarraian, 5.16. taulan sistema berdinek sortutako corpus sintetikoen aniztasun lexikala neurtzeko metriken balioak erakusten dira. Irakurterrazagoak izateko, TTR eta *Yules' I* balioak 100 aldiz biderkatuak azaltzen dira.

	Sistema	**TTR*100**	***Yules' I**100**	**MTLD**
EU	EOIA	3,70	74,30	15,33
	IAE	1,01	0,40	13,76
	LSTM	2,77	3,23	13,20
	Transformer	3,51	8,19	13,79
DE	EOIA	1,64	4,55	48,50
	IAE	0,80	0,66	74,90
	LSTM	1,90	2,31	40,00
	Transformer	2,61	5,62	53,70

5.16 taula – Atzeranzko itzulpena egiteko sistema desberdinek sortutako euskarazko eta alemanezko corpusen aniztasun lexikala neurtzeko metriken balioak

Azkenik, 5.17. taulak atzeranzko itzulpenaren bidez sortutako corpusak gehituta, eta hauei datuen hautespena aplikatu ondoren entrenatutako eu-es eta de-en sistemen IA emaitzak erakusten ditu. Hizkuntza pare bakoitzerako, lehenengo blokean atzeranzko itzulpenaren bidez sortutako corpusak gehituta entrenatutako sistemen emaitzak aurkezten dira, metrika bakoitzerako emaitza hoberenak letra lodiz markatuta; eta bigarren blokean atzeranzko itzulpenaren bidez sortutako corpusetan datuen hautespena aplikatu ondoren entrenatutako sistemen emaitzak aurkezten dira, metrika bakoitzerako emaitza hoberenak letra lodiz eta etzanez adieraziz.

Ondorioak

Corpus elebidunekin eu-es eta de-en zentzuetan entrenatutako sistemei dagokienez, 5.14. taulak erakusten duen moduan, Transformerrek emaitza askoz hobeak lortu zituen bi hizkuntza pareetan ebaluazio-metrika guztien arabera, beraz atzeranzko itzulpenaren bidez sortutako corpusa gehitu ondoren garatu beharreko sistemak entrenatzeko Transformer arkitektura erabili zen.

Kontrako zentzuetan atzeranzko itzulpena egiteko sistema desberdinen artean ere, 5.15. taulan ikusten denez, Transformer arkitekturak lortzen ditu emaitza hoberenak bi zentzuetan eta ebaluazio-metrika guztien arabera.

	Sistema	BLEU↑	TER↓	METEOR↑	chrF3↑
EU-ES	EOIA	23,27	62,67	48,02	56,51
	IAE	22,51	64,57	45,97	54,53
	LSTM	24,74	63,55	47,58	55,59
	Transformer	25,70	60,29	48,53	57,08
	GUZTIAK	**26,18**	**59,10**	**49,19**	**57,31**
	FromAll	***25,93***	59,76	48,66	56,69
	EachFromAll	25,85	***58,92***	***48,83***	***57,17***
	EachFromAll x4	24,59	61,15	48,10	56,19
	EachFromAll RS	25,77	59,86	48,59	56,92
DE-EN	EOIA	39,02	42,27	37,32	62,72
	IAE	42,32	39,21	39,37	65,91
	LSTM	40,97	39,75	38,45	64,81
	Transformer	**42,75**	38,73	39,35	**66,05**
	GUZTIAK	42,69	**38,45**	**39,65**	65,99
	FromAll	43,66	***37,71***	***40,10***	67,01
	EachFromAll	43,45	38,24	39,81	66,44
	EachFromAll RS	***43,98***	37,79	39,91	***67,10***

5.17 taula – Atzeranzko itzulpenaren bidez sortutako corpusak gehituta, eta hauei datuen hautespena aplikatu ondoren entrenatutako eu-es eta de-en sistemen IA emaitzak.

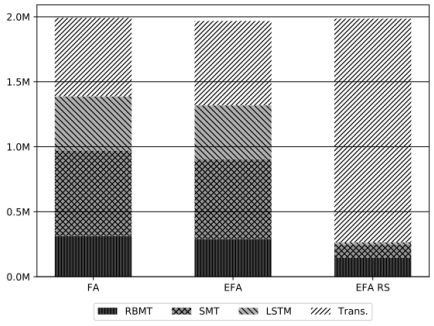
Aniztasun lexikalari dagokionez, 5.16. taulan agertzen den moduan, atzeranzko itzulpenaren bidez euskaraz sortutako corpus sintetiko desberdinen artean, EOIA bidez itzulitakoa lexikoki anitzagoa da metrika guztien arabera. Alemanezkoen artean, ordea, Transformer bidez sortutakoa anitzagoa da TTR eta *Yules' I* metriken arabera, eta IAE bidez sortutakoa MTLDren arabera.

Azkenik, 5.17. taulako emaitzak aztertuko dira. eu-es zentzuari dagokionez, ikusten dugu atzeranzko itzulpena egiteko sistema guztiek sortutako corpusak gehituta, edozein sistema bakarrik erabilita baino emaitza hobeak lortzen direla ebaluazio-metrika guztien arabera, sistema desberdinak erabiltzea aberasgarria izan daitekeela erakutsiz. Datuen hautespena aplikatzeko, *EachFromAll* hurbilpenak lortzen ditu emaitza hoberenak ebaluazio-metrika gehienen arabera, eta TER metrikaren arabera atzeranzko itzulpena egiteko sistema guztiek sortutako corpusak gehituta entrenatutako sistemaren emaitzak ere hobetzen dira, nahiz eta gehitutako corpus sintetikoa lau aldiz txi-

kiagoa izan. Aldiz, *EachFromAll x4* esperimentuan ikusten dugu atzeranzko itzulpenaren bidez sortutako corpusa 4 aldiz errepikatzea ez dela lagungarria, erabilitako corpusaren aniztasuna garrantzitsua dela iradokiz.

de-en zentzuan, atzeranzko itzulpena egiteko Transformer bakarrik erabilita edo sistema guztiak erabilita pareko emaitzak lortzen dira, horietako bakoitza bi metriken arabera gailenduz. Hau azaltzeko, 5.15. taulan atzeranzko itzulpena egiteko sistemen artean en-de zentzuan es-eu zentzuan baino desberdintasun handiagoak egotea izan daiteke arrazoietako bat. Horretaz gain, datuen hautespena edozein hurbilpenen arabera aplikatu ondoren atzeranzko itzulpen soila erabilita lortutako emaitza guztiak hobetzen dira ebaluazio-metrika guztien arabera. Gainera, proposaturiko *EachFromAll RS* birkalkulatzeko teknikak emaitza hoberenak lortzen ditu BLEU eta chrF3 metriken arabera, IA emaitzak hobetzeko aniztasun lexikala kontuan hartzeko bidea irekiz.

Datuen hautespena aplikatzeko erabilitako hurbilpen desberdinak hobeto aztertzeko, 5.1 irudiak eu-es zentzuan hurbilpen bakoitzean sistema desberdinetatik hautatutako esaldi kopuruen distribuzioa aurkezten du. Kasu honetan sistemak identifikatzeko ingelesezko akronimoak erabiltzen dira, non '*RBMT*' = EOIA, '*SMT*' = IAE; eta '*Trans.*' 'Transformer'-en laburdura den.

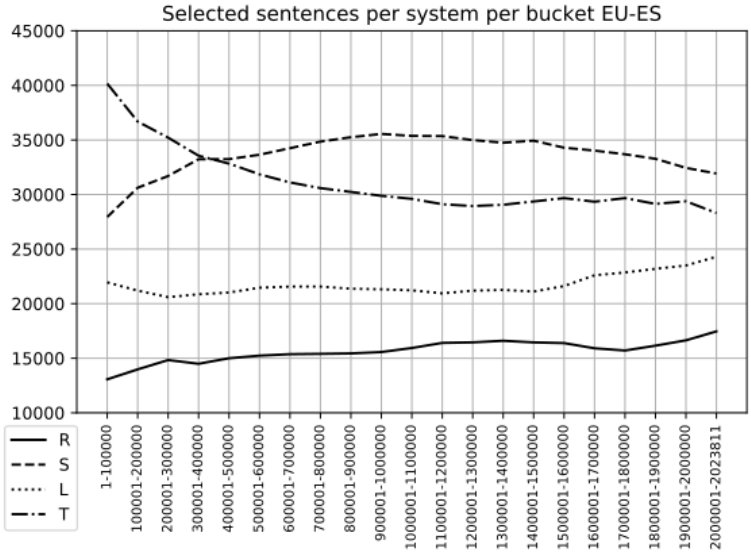


5.1 irudia – eu-es zentzuan datuen hautespena aplikatzeko erabilitako hurbilpen bakoitzarekin sistema desberdinetatik hautatutako esaldi kopuruen distribuzioa ('FA' = *FromAll*, 'EFA' = *EachFromAll*, 'EFA RS' = *EachFromAll RS*).

Ikus daitekeenez, *FromAll* eta *EachFromAll* hurbilpenetan sistema desberdinetako esaldi kopuru parekoak hautatzen dira, azkenekoak Transformer bidez itzulitako esaldi batzuk gehiago hautatuz. Honek 5.17. taulan hur-

bilpen hauen artean IA metrika gehienetan bigarrenaren aldeko alde txikia azaldu dezake. Ordea, *EachFromAll RS* hurbilpenean Transformer bidez itzulitako esaldi askoz gehiago hautatzen dira, eta honek ez du *EachFromAll* hurbilpenarekiko hobekuntzarik ekartzen. Honek erakusten du garrantzitsuena ez dela soilik sistema hoberenetik esaldi gehiago hautatzea, baizik eta aukera dagoenean jatorri desberdinetatik hautatutako esaldiak izatea aberasgarria dela.

Datuen hautespena sistema desberdinetatik sortutako corpusei aplikatzeko prozesua sakonago aztertzeko, eu-es zentzuan *FromAll* hurbilpenean sekuentzialki hautatutako 2M inguru esaldiak 100.000 esaldietako multzoetan banatu ditugu, eta multzo bakoitzean sistema bakoitzeko zenbat esaldi hautatu diren neurtu dugu. Honela, 5.2 irudian datuen hautespen prozesuak sistema desberdinetako esaldiak hautatzeko duen joera azter daiteke. Kasu honetan, sistema bakoitza identifikatzeko ingelesezko terminoaren lehenengo hizkia erabiltzen da, non 'R': EOIA, 'S': IAE, 'L': LSTM eta 'T': Transformer.



5.2 irudia – eu-es zentzuan *FromAll* hurbilpenarekin datuen hautespena aplikatzean 100.000 esaldiko multzo bakoitzean sistema desberdinetatik hautatutako esaldi kopurua (azkeneko multzorako balioak estrapolatuta)

Ikusten dugunez, hasieran Transformer bidez itzulitako esaldi gehiago

hautatzen dira, baina prozesuaren lehen erdian Transformerretik hartutako esaldien kopurua gutxitzen den heinean IAE bidez sortutako esaldi gehiago aukeratzen dira. Ordea, EOIA eta LSTM bidez sortutako esaldien kopuruak orekatuagoak mantentzen dira, azkeneko multzoetan igoera bat nabarmentzen delarik, beste bi sistemetatik hartutako esaldi kopuruak gutxitzen diren bitartean.

Atal hau amaitzeko, tesi honen helburu nagusitik haratago esperimentu multzo hauetan egindako ekarpenak laburbilduko ditugu:

1. Atzeranzko itzulpena egiteko 4 sistema desberdin erabilita sortutako corpusak batu ditugu, eu-es zentzuan sistema bakarra erabilita lortutako balioak hobetuz.
2. Lehen aldiz, atzeranzko itzulpenaren bidez sortutako corpusei datuen hautespena aplikatu diegu, hizkuntza pare batean hurbilpen guztien arabera 4 aldiz handiagoa den corpusa erabilita lortutako emaitzak hobetuz.
3. Datuen hautespenaren emaitzak atzeranzko itzulpena egiteko sistemen IA metriken eta hauen bidez sortutako corpusen aniztasun lexikalaren arabera birkalkulatzeko teknikak aztertu ditugu.

5.5 Garatutako sistemak biomedikuntzaren arloko testuak eta terminologia klinikoa ingelesetik euskarara itzultzeko moldatzea

Itzulbideko corpus elebiduna eskuragarri izan baino lehen egindako azken esperimentuetan, txosten klinikoak euskararen eta gazteleraren artean itzultzeko garatutako sistemak beste ataza batean erabiltzeko moldatu ziren: biomedikuntzaren arloko testuak ingelesetik euskarara itzultzea, testu horiek artikulu zientifikoetako laburpenetatik erauzitako esaldiak eta GNS-10eko termino klinikoak izanik. Ataza hauek WMT 2020 konferentzian biomedikuntzaren arloko testuak itzultzeko ataza partekatuan (Bawden *et al.*, 2020) definitutakoak ziren, eta atal honetan bertan izan genuen parte-hartzea des-

kribatuko dugu.

Helburua testuak ingelesetik euskarara itzultzea izanik, eta aurretik gazteleratik euskarara itzultzeko sistemak garatuak izanda, pibotatze terminoarekin ezagutzen den metodoa erabiltzea erabaki genuen, testuak ingelesetik gaztelerara itzultzeko sistemak garatuz, eta ondoren honen irteera gazteleratik euskarara itzuliz. Behin ingelesetik gaztelerara itzultzeko sistemak garatuta, corpus berdina erabilita kontrako zentzuan itzultzeko sistemak ere garatu genituen, eta sistema hauekin sortutako itzulpenak WMT 2020ko ataza partekatuan hizkuntza pare hauei zegozkien atazetara bidali genituen.

Helburua

Diseinatutako proposamenak beste ataza baterako lagungarriak izan daitezkeen aztertzea, biomedikuntzako testuak ingelesetik euskarara itzultzeko sistemak garatuz.

Corpusak

es-eu zentzurako erabilitako corpusei dagokienez, aurretik erabilitako domeinuz kanpoko 1-7 corpusei hizkuntz-identifikatzailea aplikatu zitzaien, hiztegiaren zati handiagoa bat terminologia klinikoarentzako erabili ahal izateko. Horretaz gain, SNOMED CTko terminologiaren bigarren bertsioa erabiltzen hasi zen, atazan bertan eskuragarri utzitako GNS-10 terminologiaren lehenengo bertsioa erabili zen, eta COVID-19arekin lotutako terminologia txiki batzuk erabili ziren. Gainera, akats baten ondorioz Galdakao-Usansolo Ospitaleko alta-txostenak bi aldiz batu ziren, itzulpenaren zentzua aldatuta atzeranzko itzulpena erabili ordez aurreranzko itzulpena terminoarekin ezagutzen den teknikaren bidez entrenamendu-corpusera gehituz, eta aurretik eu-es zentzuan probatutako kopiatze teknikaren bidez. Laburpen modura, 5.18. taulak es-eu sistemak entrenatzeko erabili ziren corpusak eta haien esaldi kopuruak aurkezten ditu. Barne-ebaluaziorako, aurretik erabilitako Donostiako alta-txosten ereduak erabili ziren, kasu honetan jatorrizko 2.076 esaldiak erabiliz (1.038 garapenerako eta 1.038 probarako).

Ingelesa eta gazteleraren artean itzultzeko sistemak garatzeko oinarri bezala, antolatzaileek eskuragarri utzitako Medline[8] corpusa erabili zen, 0,4M

[8] `https://github.com/biomedical-translation-corpora/corpora`

Corpusa	**Esaldiak**
Domeinuz kanpoko 1-7 corpusak (hizkuntz-identifikatzailearekin)	3.703.757
SNOMED CT 2.0 + GNS-10 1.0 + SNOMED CT / COVID-19 + Elhuyar / COVID-19 terminoak	924.804
Galdakao-Usansolo Ospitaleko gaztelerazko corpusa	2.023.811

5.18 taula – WMT 2020 konferentzian es-eu sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak

esaldi pare elebidunez osatua; eta honi garaian bildutako *TAUS Corona Crisis Corpus*-a[9] batu zitzaion, 0,9M esaldirekin. Hauekin batera, es-eu zentzuan erabilitakoen antzeko terminologia klinikoak erabili ziren (SNOMED CT, GNS-10, eta COVID-19arekin lotutakoak), kasu honetan SNOMED CT-ko terminologia ofizialetik erauzitako 385.800 termino elebidunekin. Gainera, es-en zentzuan atzeranzko itzulpena egiteko eta en-es zentzuan aurreranzko itzulpena aplikatzeko, COVID-19arekin lotutako Kaggle lehiaketa baterako bildutako CORD-19 corpusa (Wang *et al.*, 2020) erabili zen, ingelesezko 4,7M esaldiz osatua, Sketch Engine-ek prestatutako formatuan.[10] Ebaluaziorako, Khresmoi[11] erabili zen, garapenerako 500 esaldi eta probarako 1.000 esaldirekin.

Esperimentuak

Atazan 3 sistema bidaltzeko aukera izanda, ingelesaren eta gazteleraren artean itzultzeko 3 sistema desberdin garatu genituen: 1) Medline eta TAUS corpus elebidunak bakarrik erabilita; 2) aurreko corpusetara CORD-19 corpusa atzeranzko (es-en) edo aurreranzko (en-es) itzulpenaren bidez gehituta; eta 3) terminologia klinikoa gehituta. Bigarren pausoan CORD-19 corpusa gaztelerara itzultzeko lehenengo urratsean garatutako en-es sistema erabili zen.

es-eu zentzuan sistema bakarra entrenatu zen, aurretik aipatutako corpus guztiekin. Hau horrela izanda, en-eu zentzuan bidali beharreko 3 sistemak sortzeko, en-es eta es-eu sistemen *ensemble*-ak egin ziren, ikasketa-prozesuan gordetako ereduetatik garapen multzoan emaitza hoberenak lortzen zituzten 3 sistemen irteerak konbinatuz. Behin hau eginda, bidalitako en-eu sistemak

[9] `https://md.taus.net/corona`

[10] `https://www.sketchengine.eu/covid19/`

[11] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122`

honakoak izan ziren: 1) en-es sistema hoberena eta es-eu *ensemble* sistema; 2) en-es *ensemble* sistema eta es-eu sistema hoberena; eta 3) en-es *ensemble* sistema eta es-eu *ensemble* sistema.

Sistema guztiak entrenatzeko *OpenNMT*-ren Transformer inplementazioa erabili zen, defektuzko hiperparametroekin. Hizkuntza pare guztietan, atzeranzko edo aurreranzko itzulpena aplikatzerakoan dekodetze-teknika moduan *unrestricted sampling* (Edunov *et al.*, 2018) erabili zen. Sistemak ebaluatzeko BLEU metrika erabili zen (zehaztasuna ere terminologiaren itzulpena ebaluatzeko), eta antolatzaileek giza ebaluazioa egiteko zentzu bakoitzean sistema hoberena hautatzeko, sistema bakoitzarekin atazako proba multzoaren lehenengo 10/20 esaldien itzulpena eskuz aztertu zen.

Emaitzak

Lehenik eta behin, 5.19. taulan es-en, en-es eta es-eu zentzuetan gure garapen eta proba multzoetan lortutako BLEU balioak azaltzen dira. Konparaziorako, es-eu zentzuan aurreko sistema hoberenarekin lortutako balioak ere erakusten dira, 5.2. atalean domeinu klinikora egokitzeko lehenengo proben ondoren garatutako sistemari dagozkionak. Zentzu bakoitzean, letra lodiz agertzen dira emaitza hoberenak.

Noranzkoa	**Sistema**	**dev BLEU**	**test BLEU**
es-en	Medline + TAUS	56,57	52,55
	+ CORD-19 (atzeranzko itzulpena)	**61,60**	**57,25**
	+ terminologia klinikoak	60,95	56,89
en-es	Medline + TAUS	48,02	46,30
	+ CORD-19 (aurreranzko itzulpena)	**50,20**	**47,19**
	+ terminologia klinikoak	49,92	47,15
es-eu	Aurreko lana	**11,30**	**12,04**
	Lan hau	6,21	5,15

5.19 taula – WMT 2020 konferentzian es-en, en-es eta es-eu zentzuetan gure garapen eta proba multzoetan lortutako BLEU balioak, es-eu zentzuan aurreko lan hoberenaren emaitzak gehituz.

Terminologia klinikoaren eragina neurtzeko asmoz, sistema bakoitzarekin sortutako itzulpenen batez besteko luzera neurtu genuen. Balio hauek 5.20. taulan erakusten dira.

Noranzkoa	**Sistema**	**dev luzera**	**test luzera**
es-en	Medline + TAUS	20,54	22,02
	+ CORD-19 (atzeranzko itzulpena)	20,56	21,73
	+ terminologia klinikoak	20,40	21,56
en-es	Medline + TAUS	22,75	23,87
	+ CORD-19 (aurreranzko itzulpena)	22,93	23,84
	+ terminologia klinikoak	22,99	23,76

5.20 taula – WMT 2020 konferentzian es-en eta en-es zentzuetan garatutako sistemekin garapen eta proba multzoen itzulpenen batez besteko luzera. Erreferentzia bezala, garapen multzoko esaldien batez besteko luzerak 22,70 (es) eta 21,06 (en) dira; eta proba multzokoarenak 24,03 (es) eta 21,91 (en). Luzera guztiak tokenetan adieraziak dira.

Behin zentzu bakoitzerako bidali beharreko sistemak aukeratuta, hauek atazan sortutako proba multzoetan ebaluatu ziren. Lehenik eta behin, es-en, en-es eta en-eu zentzuetan laburpenak itzultzeko atazan lortutako BLEU balioak aurkezten dira 5.21. taulan. Hizkuntza pare bakoitzerako, letra lodiz aurkezten dira emaitza hoberenak, eta letra etzanez guk hoberentzat hartutakoak.

Noranzkoa	**Sistema**	**BLEU**
es-en	Medline + TAUS	*40,65*
	+ CORD-19 (atzeranzko itzulpena)	**40,71**
	+ terminologia klinikoak	39,96
en-es	Medline + TAUS	***41,71***
	+ CORD-19 (aurreranzko itzulpena)	38,36
	+ terminologia klinikoak	38,58
en-eu	(en-es) hoberena + (es-eu) *ensemble*	*8,15*
	(en-es) *ensemble* + (es-eu) hoberena	7,82
	(en-es) *ensemble* + (es-eu) *ensemble*	**8,84**

5.21 taula – WMT 2020 konferentzian atazaren proba multzoan laburpenak itzultzeko es-en, en-es eta en-eu sistemek lortutako BLEU balioak.

Azkenik, 5.22. taulan gure sistemek en-eu zentzuan terminologia klinikoa itzultzeko atazan lortutako zehaztasun eta BLEU balioak aurkezten dira.

Noranzkoa	**Sistema**	**Zehaztasuna**	**BLEU**
en-eu	(en-es) hoberena + (es-eu) *ensemble*	0,12	13,14
	(en-es) *ensemble* + (es-eu) hoberena	*0,08*	*7,21*
	(en-es) *ensemble* + (es-eu) *ensemble*	**0,13**	**14,81**

5.22 taula – WMT 2020 konferentzian en-eu zentzuan gure sistemek terminologia klinikoa itzultzeko atazaren proba multzoan lortutako zehaztasun eta BLEU balioak.

Ondorioak

Ikusten denez, gure ebaluazio-corpusen arabera, 5.19. taulan es-en eta en-es zentzuetan CORD-19 corpusa gehitu ondorengo emaitzak dira hoberenak, zentzu hauetan terminologia klinikoa gehitzea lagungarri ez zela ikusiz. Aldiz, 5.20. taulan ikusten denez, terminologia klinikoa gehitzeak sortutako itzulpenen luzera laburtzea dakar. Horretaz gain, ikusten dugu erreferentziazko garapen eta proba multzoen esaldien batez besteko luzerarekiko balio gertukoenak Medline eta TAUS corpusekin soilik entrenatutako sistemen itzulpenak direla.

es-eu zentzuari dagokionez, 5.19. taulan beherakada nabarmena ikusten da aurretik garatutako sistemaren eta lan honetarako garatutako sistemaren BLEU balioen artean. Ordea, sistema hauekin egindako atazaren proba multzoko esaldien itzulpenak aztertu ondoren, lan honetarako garatutako sistemak orohar itzulpen egokiagoak sortzen zituela ikusita, hau erabiltzea erabaki zen. Aurretik aipatu bezala, kontuan hartu behar da Donostia Unibertsitate Ospitaleko alta-txosten ereduak eta hauen gaztelerazko itzulpenak hainbat muga dituztela, kasu honetan ataza ere desberdina dela gehituz.

Edonola ere, atazan erabilitako proba multzoko esaldien itzulpenak aztertu ondoren, giza ebaluazioa egiteko eta en-eu zentzuan laburpenak eta terminologia itzultzeko Medline eta TAUS corpusekin soilik entrenatutako en-es sistema erabili genuen. Era berean, en-eu zentzuan, laburpenak itzultzeko en-es sistema hoberena eta es-eu *ensemble* sistema kateatzen zituen sistema aukeratu zen; eta terminologia itzultzeko en-es *ensemble* sistema eta es-eu sistema hoberena kateatzen zituena.

Laburpenak itzultzeko atazaren ebaluazio multzoan lortutako emaitzei dagokienez, 5.21. taulan erakusten denez, es-en zentzuan, lehenengo bi sistemek pareko emaitzak lortu zituzten; eta en-es zentzuan, hautatutako Medline + TAUS sistemak emaitza hoberenak lortu zituen nabarmen. Gainontzeko

parte-hartzaileekin alderatuta, gure sistemek emaitza baxuenak lortu zituzten es-en eta en-es zentzuetan; hala ere, egindako giza ebaluazioan gure en-es sistema lehenengo postuetan kokatu zen, nahiz eta sistema hau entrenatzeko corpus txikia erabili izan. Honek erakusten du corpusaren tamaina handia izateaz gain, honen kalitatea bermatzea ere garrantzitsua dela.

en-eu zentzuan, en-es eta es-eu *ensemble* sistemak kateatzen zituen ereduak lortu zituen emaitza hoberenak atazaren proba multzoan; eta noranzko honetan parte-hartu zuten 4 taldeetatik 2. postuan kokatu zen gure sistema.

Terminologiaren itzulpenari dagokionez, 5.22. taulan emaitza hoberenak en-es eta es-eu zentzuetan *ensemble* aukera erabilita lortu ziren, eta nahiz eta sortutako itzulpenak zentzuzkoak izan, lortutako emaitzak beste parte-hartzaileenak baino askoz baxuagoak izan ziren. Honek erakusten du terminologia klinikoa itzultzeko hobe dela corpus espezifikoak soilik erabiltzea, eta ez bestelako testuak itzultzeko garatutako sistemak berrerabiltzea.

Atal hau amaitzeko, Strubell *et al.* (2019) lanean egindako gomendioei jarraituz, WMT 2020ko atazan garatutako sistemak entrenatzeko erabilitako GPUek kontsumitutako energia eta estimatutako CO_2 emisioak aurkezten dira. Sistema hauek entrenatzeko 250Wko potentzia duten Nvidia Titan Xp GPUak erabili ziren, eta CO_2 emisioak estimatzeko Strubell *et al.* (2019) laneko (1) eta (2) ekuazioak aplikatu ziren, GPUek kontsumitutako energia bakarrik kontuan hartuta. Honela, 5.23. taulak sistema bakoitzean erabilitako GPU kopurua, entrenamendu-denbora, kontsumitutako energia eta estimatutako CO_2 emisioak erakusten ditu. Emaitzak aurkezteko, 5.19. taulan erakutsitako orden berean azaltzen dira entrenatutako sistemak.

Noranzkoa	**GPUak**	**Denbora (oo:mm)**	**Energia (kWh)**	**CO_2e (lbs)**
es-en	4	43:19	68,46	65,31
	2	46:30	36,75	35,06
	2	45:37	36,05	34,39
en-es	4	45:09	71,35	68,07
	2	47:24	37,45	35,73
	2	47:21	37,41	35,69
es-eu	2	73:14	57,86	55,20
GUZTIRA				**329,44**

5.23 taula – WMT 2020ko atazan garatutako sistemak entrenatzeko erabilitako GPU kopurua, entrenamendu-denbora, kontsumitutako energia eta estimatutako CO_2 emisioak, 5.19. taulako sistemen orden berean.

6. KAPITULUA

Metodologia eta emaitzak: domeinuko corpus elebidunarekin

Behin Itzulbideko corpus elebiduna eskura izan genuenean, eszenatokia guztiz desberdina bihurtu zen, alde batetik, corpus hau ebatzi nahi den domeinukoa izanda, egindako ebaluazioa askoz esanguratsuagoa izango delako; eta bestetik, corpus hau ebaluaziorako erabiltzeaz gain, entrenatutako sistemak birdoitzeko aukera zabaltzen delako, emaitzetan eragin nabarmena duena.

Kapitulu honetan domeinuko corpus elebidunarekin egindako esperimentuak azaldu eta beraien emaitzak aurkeztuko ditugu. Horretarako, aurreko kapituluko egitura bera jarraituko dugu, aztertutako aldagai bakoitzari atal bat eskainiz, eta horietako bakoitzean landutako helburu nagusia deskribatu, erabilitako corpusak eta sistemak aipatu, eta ateratako ondorio nagusiak zerrendatuz. Errore analisia edo giza ebaluazioa egin den kasuetan, hauek atalaren amaieran kokatu dira.

Kapitulu hau aurrekoarekin lotzeko asmoz, hasieran entrenamendu-corpus berdinarekin ebaluazio-corpus desberdinetan lortutako emaitzak azalduko dira (6.1. atala), hurrengo emaitzak hobeto interpretatzeko lagungarri izan daitekeelakoan. Ebaluazio-corpusa aldatzearekin batera, *Fairseq* erabiltzen hasi zen, beraz atal honen hasieran corpus berdinarekin *Fairseq* eta *OpenNMT* erabilita lortutako emaitzak aurkeztuko dira. Honekin batera, aurreprozesuan hainbat teknika desberdin probatzearen emaitzak azalduko dira, aurrerantzean kontuan hartu zirenak.

Behin hau eginda, Itzulbideko corpusaren 1. bertsioarekin egindako espe-

rimentu nagusien emaitzak aurkeztuko dira (6.2. atala), aldi berean Basurtoko Unibertsitate Ospitaleko gaztelerazko alta-txostenak erabiltzen hasi zirelarik. Atal honetan hitzen segmentaziorako teknika desberdinak probatuko dira, sistemak entrenamendu-corpusean agertzen ez den osasun-espezialitate batean ebaluatuko dira, eta espezialitateak entrenamendu-corpusean desberdintzeko etiketak erabiltzea probatuko da.

Ondoren, Itzulbideko corpusaren 2. bertsioarekin lortutako emaitzak aurkeztuko dira (6.3. atala), kasu honetan Basurtoko Unibertsitate Ospitaleko gaztelerazko txosten ebolutiboak ere erabiltzen hasi zirelarik. Tarte honetan atzeranzko itzulpena egiteko etiketatzea eta dekodetze-teknika desberdinak konbinatuko dira, sortutako corpus sintetikoen aniztasun lexikala aztertuz. Azken honekin lotuta, Itzulbideko corpusean dagoen genero alborapena neurtuko da; eta egindako esperimentuetan kontsumitutako energiaren estimazio bat egingo da. Dekodetze-teknika desberdinak alemanaren eta ingelesaren artean itzultzeko ere probatuko dira.

Azkenik, euskararen eta gazteleraren artean txosten klinikoak itzultzeko garatutako sistema hoberenak deskribatuko dira (6.4. atala), aurreko ataletan erabilitako corpusera Itzulbideko corpus elebakarrean datuen hautespena aplikatuz sortutako corpusa batuz. Horretarako, aurreko atalean erabilitako dekodetze-teknika hoberenak erabiliko dira, eta sortutako sistemak zein gazteleratik euskarara itzultzeko sistema ebaluazio sakon batean aztertuko dira, azken ebaluazio hau Osakidetzako langileek egin dutelarik.

6.1 Sistema aukeratu, ebaluazio-corpusa aldatu eta aurreprozesua zehaztea

Helburua

IAN sistema aldatzea, domeinuko corpus elebidunaren eragina aztertzea, eta aurrerantzean erabili beharreko aurreprozesua zehaztea.

Corpusak

Atal honen hasieran, domeinuz kanpoko 1-7 corpusak eta SNOMED CT 1.0 terminologia bakarrik erabiliko dira. Ondoren, Itzulbide 1.0 bertsioa gehituko da, honen eragina aztertzeko. Azken esperimentuetan, aurreprozesua aplikatzerakoan corpus desberdinak gehituz proba desberdinak egingo dira.

Azken esperimentu hauek deskribatzerakoan zehaztuko dira sistema bakoitza entrenatzeko erabilitako corpusak.

Esperimentuak

OpenNMT-rekin sistemak birdoitzeko arazoak izan ondoren, *Fairseq* erabiltzen hasi ginen. Aldaketa hau justifikatzeko, *OpenNMT*-rekin garatutako sistema hoberena entrenatzeko erabili zen corpus berdinarekin *Fairseq* sistema bat garatu zen, betiere Transformer arkitektura erabilita. Alderaketa hau egiteko, es-eu zentzuan egin ziren probak, gerora garatutako sistema atzeranzko itzulpena egiteko erabili ahal izateko.

Ondoren, *Fairseq*-ekin garatutako sistema Itzulbideko corpus elebidunetik erauzitako proba multzoan ebaluatu zen, honen 1. bertsiotik entrenamendurako erauzitako 24.437 esaldiak sistemak birdoitzeko erabiliz. Kontuan hartu behar da lehen proba hauetarako Itzulbideko corpus elebidunetik ebaluazio-corpusa erauzterakoan ez zela esaldi bakoitzari zegokion txosten mota kontuan hartu, beraz esaldi hauek 3.2. atalean adierazitako 5 txosten mota desberdinetakoak izan zitezkeen.

Azkenik, momentuan eskuragarri zeuden bestelako corpus batzuk gehitu ziren entrenamendu-corpusera, eta aurreprozesuan hainbat proba egin ziren. Zehazki, aurretik es-eu zentzuan erabilitako domeinuz kanpoko 1-7 corpusetara multzo bereko 8-9 corpusak batu zitzaizkien, SNOMED CT 1.0 bertsioa 2.0 bertsioarengatik ordezkatu zen, eta 4.1.2. azpiatalean aurkeztutako bestelako terminologia klinikoak gehitu ziren, GNS-10en kasuan 1. bertsioa erabiliz.

Aurreprozesuari dagokionez, egindako probak 3 izan ziren: 1) corpus osoan errepikatutako esaldiak ezabatzea; 2) kontuan hartuta domeinuz kanpoko 1. corpusa 3 aldiz errepikatuta zegoela, domeinuz kanpoko 1-7 corpusak zeuden bezala mantendu eta gainontzeko corpusetan errepikatutako esaldiak ezabatzea; eta 3) corpusak garbitzea, 100 token baino gehiago dituzten esaldiak ezabatuz.

Emaitzak

Hasteko, 6.1. taulak es-eu zentzuan aurretik *OpenNMT*-rekin garatutako sistema hoberenak eta corpus berdinarekin entrenatutako *Fairseq* sistemak Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako proba multzoan lortutako BLEU balioak aurkezten ditu.

Sistema	**test BLEU**
OpenNMT	12,04
Fairseq	**13,29**

6.1 taula – es-eu zentzuan corpus berdinarekin *OpenNMT* eta *Fairseq* tresnekin Donostia Unibertsitate Ospitaleko alta-txostenetatik erauzitako proba multzoan lortutako BLEU balioak

Jarraian, 6.2. taulan *Fairseq*-ekin garatutako sistemak Itzulbideko corpus elebidunetik erauzitako proba multzoan lortutako BLEU balioak aurkezten dira, Itzulbideko corpus elebidunaren 1. bertsioa birdoitzeko erabili aurretik (aurre-entrenamendua) eta ondoren (+ birdoitzea).

Sistema	**aurre-entrenamendua test BLEU**	**+ birdoitzea test BLEU**
Fairseq	16,57	35,29

6.2 taula – es-eu zentzuan *Fairseq*-ekin Itzulbideko corpus elebidunetik erauzitako proba multzoan lortutako BLEU balioak, ebaluazio-corpusa erauzteko esaldien txosten mota kontuan hartu gabe.

Amaitzeko, 6.3. taulak aurreprozesuaren inguruko esperimentuen emaitzak erakusten ditu, birdoitzea aplikatu aurretik eta ondoren Itzulbideko corpusetik erauzitako proba multzoan BLEUa kalkulatuz, eta erreferentzia bezala aurretik domeinuz kanpoko 1-7 corpusak eta SNOMED CT 1.0 bertsioa erabilita lortutako emaitzak gehituz. Emaitza hoberenak letra lodiz adierazten dira.

Ondorioak

Tresna desberdinei dagokienez, 6.1. taulan ikusten denez, entrenamendu-corpus berdinarekin *Fairseq* erabilita emaitzak nabarmen hobetzen dira, beraz aurrerantzean tresna hau erabiltzea erabaki zen.

Fairseq-ekin 6.1. taulan eta 6.2. taulan entrenamendu-corpus berdinarekin (aurre-entrenamendua) ebaluazio-corpus desberdinetan lortutako emaitzak alderatuta, ikusten da Itzulbideko corpus elebidunetik erauzitako proba multzoan BLEU balioak zertxobait handiagoak direla. Beraz, ondoriozta daiteke Itzulbideko corpuseko esaldiak gazteleratik euskarara itzultzea Donostia

Sistema	**aurre-entrenamendua test BLEU**	**+ birdoitzea test BLEU**
Domeinuz kanpoko 1-7 + SNOMED CT 1.0	16,57	35,29
(Domeinuz kanpoko 1-9 + terminologia klinikoak)* *errepikatutakoak kenduta	17,73	35,03
Domeinuz kanpoko 1-7 + (domeinuz kanpoko 8-9 + terminologia klinikoak)* *errepikatutakoak kenduta	18,28	35,50
Domeinuz kanpoko 1-7 + (domeinuz kanpoko 8-9 + terminologia klinikoak)* *errepikatutakoak kenduta, eta corpusa garbituta (<=100 token)	**18,39**	**35,85**

6.3 taula – es-eu zentzuan *Fairseq*-ekin aurreprozesu mota desberdinak eginez Itzulbideko corpus elebidunetik erauzitako proba multzoan lortutako BLEU balioak, ebaluazio-corpusa erauzteko esaldien txosten mota kontuan hartu gabe.

Unibertsitate Ospitaleko alta-txosten ereduak zentzu berean itzultzea baino errazagoa dela. Edonola ere, kontuan hartu behar da azken hauek jatorriz helburu akademikoekin idatziak izan direla, eta beren eskuzko itzulpenean akatsak eta itzuli gabeko terminoak ere badaudela, beraz gaztelerazko itzulpen hauek euskarara berriro itzultzean itzuli ezin diren terminoak ere egongo dira, kalkulatutako BLEU balioak murriztuz.

Bestalde, 6.2. taula ikusita argi geratzen da birdoitzeko erabilitako Itzulbideko corpus elebiduna oso baliagarria dela, ebaluazio-corpus berrian 18,72 BLEU igoz. Honek tesiaren beste ondorio nagusietako bat uzten digu:

4. Itzulbideko corpus elebiduna oso baliagarria da txosten klinikoak euskararen eta gazteleraren artean itzultzeko, emaitzak nabarmen hobetuz.

Aurreprozesuari dagokionez, 6.3. taulan ikusten dugunez, esaldi errepikatuak kenduta egindako bi esperimentuetatik (2. eta 3. lerroak) domeinuz kanpoko 1-7 corpusetan errepikatutako esaldiak mantenduta emaitza hobeak lortu ziren, beraz corpusa garbitzeko egindako proban corpus hau erabili zen. Hau eginda, entrenamendu-corpusetik 100 token baino gutxiagoko esaldiak kenduta ere emaitzak hobetzen zirela ikusi zen, beraz gainontzeko esperimentuetan hau ere kontuan hartu zen.

Hortaz, tesi honetan es-eu eta eu-es zentzuetan egindako ondorengo esperimentuetan, erabilitako corpusak edozein izanda ere, errepikatutako esaldiak ezabatu ziren (domeinuz kanpoko 1-7 corpusetan errepikatutako esaldiak mantenduz), eta ondoren 100 token baino gutxiagoko esaldiak baztertu ziren.

6.2 Itzulbide 1.0: hitzen segmentazioa, entrenamenduan ikusi gabeko espezialitateak eta hauek desberdintzeko etiketak

Helburua

Itzulbideko corpus elebidunaren lehenengo bertsioarekin, hitzen segmentaziorako teknika desberdinak probatzea, sistemak entrenamendu-corpusean agertzen ez den espezialitate batean ebaluatzea, eta entrenamendu-corpusean espezialitateak desberdintzeko etiketak erabiltzearen eragina neurtzea.

Corpusak

Esperimentu hauetarako, aurretik erabilitako domeinuz kanpoko 1-9 corpusak eta terminologia kliniko guztiak (SNOMED CT 2.0 eta GNS-10 1.0 bertsioak barnebilduz) erabili ziren. Hauei, E3C proiekturako (Magnini *et al.*, 2020) bildutako Basurtoko Unibertsitate Ospitaleko saio klinikoetatik erauzitako 541 esaldi pare elebidun gehitu zitzaizkien, Itzulbideko corpus elebidunaren 1. bertsioarekin batera sistemak birdoitzeko erabili zirenak. Horretaz gain, atzeranzko itzulpena eta kopiatze teknika aplikatzeko, aurretik erabilitako Galdakao-Usansolo Ospitaleko alta-txostenen bertsio murriztuaz gain, Basurtoko Unibertsitate Ospitaleko alta-txostenak ere erabili ziren. Laburpen modura, 6.4. taulak atal honetan deskribatutako sistemak entrenatzeko erabili ziren corpusak eta haien esaldi kopuruak aurkezten ditu.

Corpusa	**Esaldiak**
Domeinuz kanpoko 1-9 corpusak	4.292.546
SNOMED CT 2.0 + GNS-10 1.0 + SNOMED CT / COVID-19 + Elhuyar / COVID-19 terminoak	924.804
Itzulbide 1.0 + Basurtoko Unibertsitate Ospitaleko saio klinikoak	24.978
Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko alta-txostenak	2.827.565

6.4 taula – Itzulbideko corpus elebidunaren lehenengo bertsioarekin garatutako sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak

Esperimentuak

Egindako lehenengo esperimentu multzoan, es-eu zentzuan hitzen segmentaziorako algoritmo desberdinak probatu ziren, gure atazarako emaitza hoberenak lortzen zituen teknika identifikatzeko asmoz. Modu honetan, orain arte erabilitako BPE teknikaz gain, BPE-dropout probatu zen, honen bi aldaeretan: 1) corpus elebidunean aplikatuta; eta 2) jatorrizko hizkuntzan bakarrik aplikatuta. Izan ere, Provilkov *et al.* (2020) lanean adierazten den moduan, entrenamendu-corpusaren tamainaren arabera emaitzak hobeak izan daitezke aldaera hauetako bakoitzean, eta guk erabilitako corpusak lan horretan probatutako tamainen erdibideko balio bat dauka.

Teknika hauek bi modutara ebaluatu ziren; horretarako, lehenik eta behin Itzulbideko corpus elebidunaren 1. bertsiotik traumatologia espezialitateko esaldi guztiak erauzi ziren, eta amaitutzat emandako dokumentuetatik zetozenak traumatologiako proba multzo berezitua osatzeko erabili ziren. Ondoren, traumatologiakoak ez ziren espezialitateetako esaldietatik ausaz erauzi ziren 1.000 esaldi garapenerako eta beste 1.000 probarako. Horrela, hitzen segmentaziorako algoritmo desberdinen emaitzak aztertzeaz gain, gure sistemek entrenamendu-corpusean agertzen ez den espezialitate batean lortzen dituzten emaitzak aztertu nahi izan genituen. Traumatologia espezialitatea aukeratu genuen esaldi kopurua proba multzokoaren parekoa izateagatik, eta terminologiaren ikuspuntutik gainontzeko espezialitateekiko desberdina izateagatik.

es-eu zentzuan egindako bigarren esperimentuan Itzulbideko corpusaren 1. bertsiotik erauzitako esaldiei etiketa bat gehitu zitzaien bakoitzaren espezialitatea adierazteko. Modu honetan, sistemak esaldi bakoitzaren espe-

zialitatea ezagutzeak itzulpenen kalitatea hobetzen zuen ala ez aztertu nahi genuen. Hau egiteko, 3.3. taulan agertzen diren espezialitate bakoitzarentzat 3 karakterezko akronimo bat definitu genuen, '<' eta '>' sinboloen artean txertatuz (adib.: '<DIG>' "digestio-aparatua" espezialitaterako). Behin etiketa hauek definituta, Itzulbideko corpusaren 1. bertsioko esaldi bakoitzaren hasieran gehitu genuen, ondoren zuriune bat utziz. Etiketa hauek hitzen segmentazioa aplikatu ondoren gehitu genituen, ikasitako BPE ereduan eraginik izan ez zezaten.

Sistema hauek ebaluatzeko aurrerantzean erabiliko zen ebaluazio-corpusa erabiltzen hasi ginen, espezialitaterik baztertu gabe alta-txostenak eta txosten ebolutiboak bakarrik kontuan hartuz erauzitako 2.000 esaldiez osatua, 1.000 garapenerako eta beste 1.000 probarako erabiliz.

eu-es zentzurako bi esperimentu egin genituen, es-eu zentzuan erabilitako corpusera Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko alta-txostenak atzeranzko itzulpenaren bidez gehituta, eta ondoren kopiatze teknikaren bidez gehituta. Corpus elebakarren eragina hobeto aztertzeko, sistemak Itzulbideko corpusarekin birdoitzea aplikatu aurretik eta ondoren ebaluatuko dira.

Emaitzak

Hasteko, 6.5. taulak es-eu zentzuan hitzen segmentaziorako algoritmo desberdinak erabilita Itzulbideko corpusaren 1. bertsiotik erauzitako proba multzo orokorrean eta traumatologiako proba multzoan lortutako BLEU balioak aurkezten ditu. BPE aplikatzerakoan sistemak 20 *epoch*-etan aurre-entrenatu eta birdoitu ziren, eta BPE-dropout aplikatzerakoan prozesu berdina 50 *epoch*-etan egin zen. Balio hoberenak letra lodiz adierazten dira.

Sistema	**orokorra test BLEU**	**traumatologia test BLEU**
BPE	36,24	19,28
BPE-dropout (elebitan)	**37,42**	**19,52**
BPE-dropout (jatorrian)	37,06	18,57

6.5 taula – es-eu zentzuan hitzen segmentaziorako algoritmo desberdinak erabilita Itzulbideko corpusaren 1. bertsiotik erauzitako proba multzo orokorrean eta traumatologiako proba multzoan lortutako BLEU balioak

Ondoren, 6.6. taulak es-eu zentzuan alta-txosten eta txosten ebolutiboetatik erauzitako proba multzoan lortutako BLEU balioak aurkezten dira, espezialitateak identifikatzeko etiketak gehituta eta gehitu gabe. Emaitza hoberena letra lodiz adierazten da.

Sistema	**alta-txostenak eta txosten ebolutiboak test BLEU**
Etiketarik gabe	**31,97**
Etiketekin	30,67

6.6 taula – es-eu zentzuan espezialitateak identifikatzeko etiketak gehituta eta gehitu gabe Itzulbideko corpusaren 1. bertsiotik alta-txostenak eta txosten ebolutiboak bakarrik kontuan hartuta erauzitako proba multzoan lortutako BLEU balioak

Azkenik, 6.7. taulak eu-es zentzuan Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko alta-txostenak atzeranzko itzulpenaren bidez gehituta, eta ondoren kopiatze teknikaren bidez gehituta lortzen diren BLEU balioak aurkezten ditu. Kasu honetan ebaluazio-corpusa es-eu zentzuan azkeneko esperimentuak ebaluatzeko erabili zena izan zen, Itzulbideko corpusaren 1. bertsiotik alta-txostenak eta txosten ebolutiboak bakarrik kontuan hartuta erauzitakoa. Emaitza hoberenak letra lodiz adierazten dira.

Sistema	**aurre-entrenamendua test BLEU**	**+ birdoitzea test BLEU**
Atzeranzko itzulpena	38,50	**50,67**
+ kopiatzea	**38,74**	49,64

6.7 taula – eu-es zentzuan Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko alta-txostenak atzeranzko itzulpenaren bidez gehituta, eta ondoren kopiatze teknikaren bidez gehituta Itzulbideko corpusaren 1. bertsiotik alta-txostenak eta txosten ebolutiboak bakarrik kontuan hartuta erauzitako proba multzoan lortutako BLEU balioak.

Hitzen segmentazioari dagokionez, 6.5. taulan ikusten denez, BPE-dropout bi hizkuntzetan aplikatuta lortzen dira emaitza hoberenak, beraz hurrengo esperimentuetarako teknika hau erabili zen. Honek esperimentu multzo hauetatik ateratako tesiaren beste ondorio bat ematen digu:

5. Txosten klinikoak euskararen eta gazteleraren artean itzultzeko, hitzen segmentaziorako BPE-dropout teknikak lortzen ditu emaitza hoberenak, corpusaren bi hizkuntzetan aplikatuz.

Bestalde, 6.5. taulan traumatologiako proba multzoko emaitzak ikusita pentsatu genezake sistemaren kalitatea nabarmen okertzen dela entrenamendu-corpusean agertzen ez den espezialitate bateko esaldietan ebaluatzerakoan, baina sortutako itzulpenak aztertuta bestelako ondorio batetara iristen gara. Izan ere, itzulpen automatikoen eta erreferentziazko eskuzko itzulpenen arteko desberdintasun gehienak eskuzko itzulpenen akatsei dagozkie. Desberdintasun hauek hainbat kategoriatan sailka daitezke, hala nola akats tipografikoak (‘egunena’ ‘egunean’-en ordez), akats ortografikoak (‘protezi’ ‘protesi’-ren ordez) edota euskalki desberdinak erabiltzearen ondorioak (’barik’ ’gabe’-ren ordez). Aztertutako 100 esaldietan ez zen traumatologia espezialitateari bereziki egokitutako akatsik topatu; edonola ere, gure sistema orain arte jasotako corpusean agertzen ez den espezialitate bateko esaldiak itzultzeko erabili baino lehen osasun-langileek ebaluatzea gomendatzen dugu.

Etiketen eragina aztertzeko orduan, 6.6. taulan ikus daitekeenez, emaitza hoberenak etiketarik gehitu gabe lortzen dira. Horren arrazoietako bat 3.3. taulan espezialitate bakoitzeko esaldi kopurua oso desberdina izatea izan daiteke, etiketak gehitzearen eragina murriztuz. Edonola ere, emaitza hauek kontuan hartuta, atzeranzko itzulpena egiteko etiketarik gabe entrenatutako sistema erabili zen, eta gainontzeko esperimentuetan ez zen espezialitateak bereizteko etiketarik erabili.

eu-es zentzuari dagokionean, 6.7. taulan ikusten denez, behin Itzulbideko corpusa gehituta eta corpus elebakarra handiagoa izanda, kopiatze teknikak jada ez du aurretik 5.6. taulan ikusitako hobekuntzarik ekartzen, beraz aurrerantzean corpus elebakarrak atzeranzko itzulpenaren bidez soilik gehitu ziren entrenamendu-corpusera.

Erroreen analisia

Atal hau amaitzeko, zentzu bakoitzean garatutako sistema hoberenek osasun-txostenak itzultzeko garrantzi berezia duten termino batzuk nola itzultzen diren aztertuko dugu.

Hasteko, osasun-txostenetan agertzen diren zenbakiak (adib.: analitiketa-

ko emaitzak) gaizki itzultzeak izan lezakeen eragina kontuan hartuta, proba multzoko lehen 100 lerroetan agertzen ziren zenbakiak nola itzuli ziren aztertu genuen, eta baieztatu genuen bi zentzuetan zenbaki guztiak zuzentasunez itzultzen zirela.

Azterketa sinple honen ondoren, terminologia klinikoa nola itzultzen zen aztertu genuen. Proba multzoko esaldietan agertzen zen terminologia klinikoa identifikatzeko, SNOMED CT 2.0 bertsioan agertzen diren 896.898 termino elebidunak bilatu genituen proba multzoko corpusean, eta zegozkien esaldietan gure sistemak termino horiek nola itzultzen zituen aztertu genuen. Azterketa hau es-eu zentzuan egin genuen. SNOMED CTko terminologia automatikoki euskaratua izanda, bilaketa hau terminoen sorkuntza egiteko erabili ziren metodoen arabera baldintzatuta dago. Modu honetan, proba multzoan topatu ziren SNOMED CTko termino elebidun guztiak bi hizkuntzetan berdin idazten ziren edo bata bestearen transliterazioa ziren, SNOMED CT euskaratzeko erabilitako metodoetako bat transliteratzea izanik. Zehazki, proba multzoko 1.000 esaldietatik 37 esaldietan identifikatu ziren SNOMED CTko terminoak, guztira 44 termino aurkituz. Termino hauek guztiak modu egokian itzuliak izan ziren. Sortutako itzulpenetan agertzen ziren terminoen eta automatikoki euskaratutako SNOMED CTko terminoen arteko desberdintasun bakarrak 2 izan ziren: 1) 'elektrokardiograma'-ren ordez bere akronimoa den 'EKG' itzultzea; eta 2) 'disnea'-ren ordez 'arnashestu' edo 'arnasestu' terminoak itzultzea. Azken honi dagokionez, nahiz eta 'arnas-hestu' edo 'arnasestu' terminoak SNOMED CTren euskaratzean ez agertu, ontzat har genitzake, entrenamendu-corpusean agertzeaz gain, ingelesezko SNOMED CTn 'disnea'-ren sinonimo moduan agertzen den '*breathless*' terminoaren parekoak izateagatik.

Aurreko azterketa osatzeko, eu-es zentzuan ere SNOMED CTko terminoak bilatu ziren proba multzoan, kasu honetan ere termino guztiak modu zuzenean itzuliak izanik. Ordea, SNOMED CTko terminoak jatorrizko euskarazko corpusean bakarrik bilatzean, 'tartsorrafia' terminoa modu okerrean itzultzen zela ikusi zen, erreferentziazko 'tarsorrafia' terminoaren ordez 'tarstorafia' itzuliz. Edonola ere, SNOMED CT automatikoki euskaratua izatearen mugak kontuan izanik, honelako azterketak modu egokian egin ahal izateko komenigarria litzateke euskarazko terminologia klinikoa ofiziala izatea.

Muga hauek kontuan hartuta, biomedikuntzan aditu bati eu-es zentzuan proba multzotik ausaz erauzitako 100 esaldietan termino klinikoak eskuz identifikatzeko eskatu genion. Modu honetan, 'neumokoko polisakaridoa

23Varen kontrako txertoa' terminoa 'vacuna polisacarida de neumococo 23V' bezala itzultzen zela ikusi genuen, erreferentziazko 'vacuna antineumocócica polisacárida 23V' zuzenaren ordez.

6.3 Itzulbide 2.0: atzeranzko itzulpenerako dekodetze-teknika desberdinak probatu eta sortutako corpusen aniztasun lexikala aztertzea

Helburua

Atzeranzko itzulpena egiteko etiketatzea eta dekodetze-teknika desberdinak konbinatzearen eragina ikertzea, sortutako corpus sintetikoen aniztasun lexikala aztertuz. Azken honen barruan, domeinuko corpus elebidunaren genero alborapena neurtzea.

Corpusak

Esperimentu hauetarako, aurretik erabilitako domeinuz kanpoko 1-9 corpusetara EiTB (2020) corpusa gehitu genion, GNS-10 1.0 bertsioa 2.0 bertsioarengatik ordezkatu genuen, Itzulbideko corpusaren 2. bertsioa erabili genuen, eta Basurtoko Unibertsitate Ospitaleko gaztelerazko txosten ebolutiboak ere atzeranzko itzulpenerako erabili genituen. Hau da, 4.1. atalean aurkeztu eta sistemak entrenatzeko erabili ziren baliabide desberdinetatik, Itzulbideko corpus elebakarra izan ezik corpus guztiak erabili genituen, corpus bakoitzaren bertsio bat baino gehiago zegoenean azkeneko bertsioa erabiliz. Laburpen moduan, 6.8. taulak atal honetan deskribatutako sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak aurkezten ditu. Ebaluaziorako Itzulbideko corpus elebidunetik alta-txostenak eta txosten ebolutiboak bakarrik kontuan hartuta erauzitako garapen eta proba multzoak erabili ziren, Donostia Unibertsitate Ospitaleko alta-txosten ereduak baztertuz.

Gainera, corpus pribatuekin egindako esperimentuak beste hizkuntza pare batean erreproduzitu ahal izateko, esperimentu berdinak alemanaren eta ingelesaren artean errepikatu genituen. Horretarako, corpus elebidun moduan WMT 2020 konferentzian biomedikuntzaren arloko testuak itzultzeko

Corpusa	**Esaldiak**
Domeinuz kanpoko 1-10 corpusak	4.929.728
SNOMED CT 2.0 + GNS-10 2.0 + SNOMED CT / COVID-19 + Elhuyar / COVID-19 terminoak	926.778
Itzulbide 2.0 + Basurtoko Unibertsitate Ospitaleko saio klinikoak	29.346
Galdakao-Usansolo Ospitaleko alta-txostenak, Basurtoko Unibertsitate Ospitaleko alta-txostenak eta Basurtoko Unibertsitate Ospitaleko txosten ebolutiboak	5.145.926

6.8 taula – Atzeranko itzulpenerako dekodetze-teknika desberdinak alderatzeko garatutako sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak

ataza partekatuan (Bawden *et al.*, 2020) oinarri-lerro sistemak entrenatzeko erabilitako UFAL corpusa[1] erabili genuen, "*Subtitles*" azpimultzoa erauzi ondoren 3M esaldi inguruz osatua. de-en hizkuntza parean atzeranzko itzulpena aplikatzeko Mimic III (Johnson *et al.*, 2016) corpuseko ingelesezko alta-txostenak erabili genituen.[2] Lerro hutsak eta dokumentu bakoitzaren goiburuan agertzen ziren metadatuak ezabatu ondoren, 2M esaldi inguru geratu ziren erabilgarri. Ebaluaziorako Khresmoi[3] erabili genuen, Bawden *et al.* (2020) lanean ere erabilitakoa, non 500 esaldi garapenerako eta 1.000 probarako erabiltzen diren. Bai ebaluaziorako baita *beam search* bidez corpus elebakarra itzultzeko, erabilitako *beam-width*-a 16 izan zen. Balio hau, BPE prozesuaren 40.000 iterazioak bezala, en-de zentzurako optimizatuak izan ziren Bawden *et al.* (2020) lanean.

Esperimentuak

Tesi honetan egindako azken esperimentu multzo nagusi honetan, atzeranzko itzulpena egiterakoan erabili daitezkeen dekodetze-teknika nagusiak aztertzea izan zen helburua. Modu honetan, aurreko esperimentu desberdinetan erabilitako *beam search* (Tillmann eta Ney, 2003) eta *unrestricted sampling* (Edunov *et al.*, 2018) metodoekin batera, azken honen ondoren aurkeztutako *restricted sampling* (Graça *et al.*, 2019) metodoa ere probatu genuen. Teknika hau funtsean *unrestricted sampling* metodoaren egokitzapen bat da, sortu

[1] `https://ufal.mff.cuni.cz/ufal_medical_corpus`

[2] `https://mimic.physionet.org/gettingstarted/access/`

[3] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122`

beharreko hurrengo hitza ausaz hautatzerakoan hitz posibleak probabilitate altuena duten n hitzei mugatuz, edota probabilitate minimo bat gainditzen duten hitzak bakarrik hautatuz. Bestalde, azken artikulu hau aurkeztu zen konferentzia berean, dekodetze-teknika aldatu ordez atzeranzko itzulpenaren bidez sortutako corpusa etiketatzea proposatu zen (Caswell *et al.*, 2019), corpus sintetikoaren esaldi bakoitzaren hasieran '<BT> ' etiketa gehituz, entrenatutako sistemak corpus elebiduna eta sintetikoa desberdintzeko gai izaten laguntzeko asmoz. Ordea, etiketatze hau *beam search* bidez eta Edunov *et al.* (2018) lanean aurkeztutako *noising* metodoaren bidez sortutako corpusei aplikatu zitzaien, baina ez lan berean aurkeztutako *unrestricted sampling* edo aldi berean aurkeztutako *restricted sampling* tekniken bidez sortutako corpusei.

Hortaz, gure helburua etiketatzea aipatutako dekodetze-teknika desberdin guztiekin konbinatzea izango da. Horretarako, atal honetan atzeranzko itzulpenaren bidez corpusa sortzeko 6 modu desberdin alderatuko ditugu, baldin eta dekodetzeko *beam search*, *restricted sampling* edo *unrestricted sampling* erabili zen; eta horietako bakoitzaren bidez sortutako corpusa etiketatu zen edo ez. Esperimentu hauek konparagarriak izateko eta sistemen entrenamendu-denborak murrizteko, BPE-dropout erabili ordez ohiko BPE erabili zen hitzen segmentaziorako, ondoren emaitza hoberenak lortzen zituzten sistemak BPE-dropout bidez aurreprozesatzeko asmoz.

eu-es zentzuko sistema hoberenak zeintzuk ziren argitzeko giza ebaluazio murriztu bat egin genuen, IA metrika altuenak lortzen zituzten 3 sistemak alderatuz. Giza ebaluazio hau egiteko, biomedikuntzan aditu elebidun bati eu-es zentzuan IA metrika altuenak lortu zituzten sistemen itzulpenak modu itsuan ebaluatzeko eskatu genion. Horretarako, itzulpenen zuzentasunean jarri genuen fokua, esaldi bakoitzeko jatorrizko testuaren eta sistema bakoitzaren itzulpenaren esanahia alderatuz. Lagin moduan, proba multzoko lehen 100 esaldi ez-errepikatuak hartu genituen, eta sistema bakoitzak modu guztiz zuzenean zenbat esaldi itzultzen zituen zenbatu.

Bestalde, aurretik 5.4. atalean egin bezala, atzeranzko itzulpenaren bidez sortutako corpus desberdinen aniztasun lexikala aztertu zen, corpus hauekin garatutako sistemen kalitatean eraginik ote zuen aztertzeko asmoz. Ebaluaziorako, 2. kapituluan aurkeztutako 4 IA metrikak kalkulatu ziren. Horretaz gain, Itzulbideko corpus elebidunean termino estereotipatuen genero alborapena neurtu zen, eta garatutako sistemek sortutako CO_2-a estimatu zen.

Emaitzak

Hasteko, erreferentzia bezala, 6.9. taulak es-eu eta en-de zentzuetan entrenatutako sistemen IA metriken emaitzak erakusten ditu, gerora atzeranzko itzulpena egiteko erabili zirenak.

Noranzkoa	**BLEU↑**	**TER↓**	**METEOR↑**	**CHRF↑**
es-eu	33,88	49,27	47,02	61,02
en-de	29,96	52,63	47,64	60,60

6.9 taula – Atzeranzko itzulpena dekodetze-teknika desberdinekin aplikatzeko es-eu eta en-de zentzuetan entrenatutako sistemen IA emaitzak

Jarraian, 6.10. taulan eu-es eta de-en zentzuetan entrenatutako sistemen IA emaitzak aurkezten dira. Hizkuntza pare bakoitzerako, emaitzen lehen lerroan atzeranzko itzulpenaren bidez sortutako corpusa gehitu aurretik lortutako emaitzak aurkezten dira (aurre-entrenamendua); eta eu-es zentzuan emaitzak birdoitu aurretik eta ondoren erakusten dira. Dekodetze-teknika desberdinak adierazteko ingelesezko terminoak erabiltzen dira, '*unr.*' '*unrestricted*'-en laburdura izanda eta '*res.*' '*restricted*'-ena. Bestalde, etiketatzea adierazteko ingelesezko '*tagged*' terminoaren '*tag.*' laburdura erabiltzen da, eta atzeranzko itzulpenari erreferentzia egiteko ingelesezko '*back-translation*' terminoaren '*BT*' akronimoa erabiltzen da. Emaitza hoberenak letra lodiz adierazten dira.

Bestalde, 6.11. taulak atzeranzko itzulpena egiteko dekodetze-teknika desberdinak erabiliz sortutako euskarazko eta alemanezko corpusen aniztasun lexikala neurtzeko metriken balioak erakusten ditu, corpus sintetikoa etiketatuz edo ez. TTR eta *Yules' I* balioak 100 aldiz biderkatuak azaltzen dira, irakurterrazagoak izateko.

Azkenik, metrika automatikoetan balio altuenak lortu zituzten sistemak (*tagged restricted sampling*, *restricted sampling* eta *unrestricted sampling*) giza ebaluazio murriztu batean aztertu ziren. Horrela, 6.12. taulak eu-es zentzuan IA metrika altuenak lortu zituzten sistemek proba multzoko lehen 100 esaldi ez-errepikatuetatik zenbat esaldi modu guztiz zuzenean itzultzen zituzten erakusten du.

	Sistema	BLEU↑	TER↓	METEOR↑	CHRF↑
EU-ES	aurre-entrenamendua	26,99	58,61	47,70	53,35
	+ birdoitzea	46,67	38,74	63,56	66,46
	+ *BT (beam search)*	44,11	41,54	61,48	66,24
	+ birdoitzea	51,37	35,15	67,11	70,10
	+ *BT (tag. beam search)*	41,29	44,45	59,47	64,22
	+ birdoitzea	51,99	34,96	67,27	70,11
	+ *BT (unr. sampling)*	43,48	41,39	61,36	65,94
	+ birdoitzea	52,68	33,84	67,93	71,06
	+ *BT (tag. unr. sampling)*	42,07	44,33	59,97	65,13
	+ birdoitzea	52,42	34,75	67,51	70,72
	+ *BT (res. sampling)*	44,69	40,83	62,23	66,85
	+ birdoitzea	52,90	33,96	68,23	71,12
	+ *BT (tag. res. sampling)*	42,13	43,71	60,22	65,40
	+ birdoitzea	**53,10**	**33,55**	**68,30**	**71,34**
DE-EN	aurre-entrenamendua	42,34	38,55	39,91	67,93
	+ *BT (beam search)*	**44,67**	**37,46**	**40,97**	**69,62**
	+ *BT (tag. beam search)*	44,40	37,63	40,79	69,41
	+ *BT (unr. sampling)*	42,47	41,17	39,58	67,65
	+ *BT (tag. unr. sampling)*	43,14	38,42	40,35	68,59
	+ *BT (res. sampling)*	40,03	45,73	38,60	66,42
	+ *BT (tag. res. sampling)*	43,27	38,28	40,51	68,68

6.10 taula – Atzeranzko itzulpena egiteko dekodetze-teknika desberdinak erabiliz eu-es eta de-en zentzuetan entrenatutako sistemen IA emaitzak, corpus sintetikoa etiketatuz edo ez.

Ondorioak

Atzeranzko itzulpena egiteko dekodetze-teknika desberdinak etiketatzearekin konbinatuta lortutako emaitzak alderatuta, 6.10. taulan ikusten dugunez, eu-es zentzuan birdoitzea aplikatu ondoren entrenatutako sistemen artean proposaturiko konbinazioetako batek ('*tag. res. sampling*': *restricted sampling* bidez sortutako corpusa etiketatuz gehitzean) lortzen ditu emaitza hoberenak ebaluazio-metrika guztien arabera, *restricted sampling* eta *unrestricted sampling* metodoez jarraitua.

de-en zentzuan, non erabilitako corpusaren tamaina txikiagoa den, ohiko *beam search* teknikarekin lortzen dira emaitza hoberenak, corpus sintetiko berdina etiketatu ondoren (*tag. beam search*) gehituta emaitza parekoak

Hizkuntza	**Corpusa**	**MTLD**	***Yule's I**100**	**TTR*100**
eu	*BT (beam search)*	13,71	0,863	0,578
	BT (tag. beam search)	14,72	0,799	0,387
	BT (unr. sampling)	13,99	7,628	65,22
	BT (tag. unr. sampling)	14,84	7,123	41,69
	BT (res. sampling)	13,73	2,545	5,851
	BT (tag. res. sampling)	14,72	2,359	3,748
de	*BT (beam search)*	14,50	0,899	0,754
	BT (tag. beam search)	15,37	0,841	0,521
	BT (unr. sampling)	15,15	8,376	93,62
	BT (tag. unr. sampling)	15,86	7,890	62,19
	BT (res. sampling)	14,39	3,374	12,64
	BT (tag. res. sampling)	15,15	3,167	8,566

6.11 taula – Atzeranzko itzulpena egiteko dekodetze-teknika desberdinak erabiliz sortutako euskarazko eta alemanezko corpusen aniztasun lexikala neurtzeko metriken balioak, corpus sintetikoa etiketatuz edo ez.

tagged restricted sampling	***restricted sampling***	***unrestricted sampling***
83	75	83

6.12 taula – eu-es zentzuan IA metrika altuenak lortu zituzten sistemek proba multzoko lehen 100 esaldi ez-errepikatuetatik modu guztiz zuzenean itzulitako esaldi kopurua

lortuz. Hizkuntza pare honetan, ikusten dugu dekodetze-teknika '*sampling*' denean, corpus sintetikoa etiketatzeak emaitzak hobetzen dituela, Caswell *et al.* (2019) lanean *beam search* bidez sortutako corpusak etiketatuz lortutako emaitzak osatuz eta '*sampling*' erabiltzea eta corpus sintetikoak etiketatzea osagarriak direla erakutsiz.

Aniztasun lexikalari dagokionez, 6.11. taulan hizkuntza bakoitzean lortutako balioak alderatuz, harrigarria suertatzen da corpus sintetikoaren esaldi guztiei etiketa berdina gehitzean MTLD balioak igo egiten direla. Hau alde batera utzita, ikusten dugu etiketatu gabeko corpusen artean *unrestricted sampling* bidez itzulitakoan lortzen direla aniztasun lexikala neurtzeko metriken balio altuenak. *Restricted sampling*-ek, ordea, nahiz eta aniztasun lexikala neurtzeko metriketan balio baxuagoak lortu, 6.10. taulan IA metrika altuagoak lortu zituen eu-es zentzuan.

Giza ebaluazio murriztuaren emaitzak aztertzerakoan, 6.12. taulan ikusten denez, *tagged restricted sampling* eta *unrestricted sampling* metodoekin lortzen dira emaitza hoberenak eu-es zentzuan, 100 esaldietatik 83 modu guztiz zuzenean itzuliz. Ordea, *restricted sampling* metodoak, 6.10. taulan IA metrika gehienen arabera 2. emaitza hoberenak lortzen zituenak, beste bi sistemek baino nabarmen okerrago itzultzen du, 100 esaldietatik 75 zuzenki itzuliz. Kontuan hartuta sistema honek 6.11. taulan MTLD metrikaren arabera balio baxuenak lortzen zituela, ondorioztatu dezakegu IA sistemen kalitatea ebaluatzerakoan IA metrikei begiratzeaz gain, aniztasun lexikala neurtzeko metrikak ere garrantzitsuak direla sistema hoberenak hautatzerakoan. Giza ebaluazio honen emaitzak kontuan hartuta, hurrengo atalean azken sistemak garatzeko *tagged restricted sampling* eta *unrestricted sampling* metodoak erabiliko dira dekodetze-teknika moduan.

Aniztasun lexikalak duen garrantziaren adibide moduan, euskaraz generorik ez duten 'paziente', 'erizain' eta 'mediku' terminoak gazteleraz zein formatan agertzen diren aztertu genuen. Horretarako, Itzulbideko corpus elebidunaren 2. bertsioa erabili genuen aztergai moduan, entrenamendu, garapen eta proba multzoetan zenbatutako agerpenak desberdinduz. Modu honetan, 6.13. taulak Itzulbideko corpus elebidunaren gaztelerazko zatian '*el paciente*', '*la paciente*', '*el enfermero*', '*la enfermera*', '*el médico*' eta '*la médico*' testuen agerpenak erakusten ditu.[4]

Terminoa	**entrenamendua**	**garapena**	**proba**
el paciente	129	1	4
la paciente	120	6	5
el enfermero	1	0	0
la enfermera	10	0	0
el médico	65	0	0
la médico	5	0	0

6.13 taula – Itzulbideko corpus elebidunaren gaztelerazko zatian 'paziente', 'erizain' eta 'mediku' terminoen gaztelerazko forma ohikoenen agerpenak entrenamendu, garapen eta proba multzoetan.

[4]Nahiz eta 'mediku' terminoaren gaztelerazko forma femenino onartua '*médica*' izan, termino hau bitan bakarrik agertzen zen (behin entrenamendu-corpusean eta beste behin garapen corpusean), beraz gehiagotan agertzen den '*la médico*' testua lehenetsi genuen.

Ikusten dugunez, estereotiporik gabeko 'paziente' terminoa modu parekoan agertzen da Itzulbideko corpus elebidunaren gaztelerazko zatian bere forma maskulino eta femeninoan. Kontrara, estereotipatutako 'erizain' eta 'mediku' terminoak hainbat aldiz gehiago agertzen dira gaztelerazko forma estereotipatuetan. 'Mediku' terminoarena bereziki harrigarria da, kontuan hartuta Osakidetzako mediku gehienak emakumeak direla. Honen arrazoi moduan, gure hipotesia da baimen informatuetan forma maskulinoa gehiago erabiltzen dela mediku zehatz bati erreferentzia egiten ez zaionean. Edonola ere, ikusita 'erizain' eta 'mediku' terminoak ez direla garapen eta proba multzoetan agertzen, ezin dugu gure sistemen genero alborapena neurtu. Hau kontuan hartuta, etorkizuneko lan bezala uzten dugu euskararen eta gaztelararen artean domeinu klinikoko testuetan genero alborapena neurtzeko proba multzo espezifikoa definitzea.

Esperimentu multzo honen azterketa amaitzeko, sistemak garatzeko erabili genituen GPUek kontsumitutako energia eta estimatutako CO_2 emisioak erakutsiko ditugu. Aurretik egin moduan, sistema hauek entrenatzeko Nvidia Titan Xp GPUak erabili ziren, 250Wko potentziarekin; eta CO_2 emisioak estimatzeko Strubell *et al.* (2019) laneko (1) eta (2) ekuazioak aplikatu ziren, GPUek kontsumitutako energia bakarrik kontuan hartuta. Modu honetan, 6.14. taulak sistema bakoitzari dagozkion entrenamendu-denbora, kontsumitutako energia eta estimatutako CO_2 emisioak erakusten ditu. Kontuan hartu behar da es-eu sistema garatzeko BPE-dropout erabili zela, entrenamendua 50 *epoch*-ez luzatuz; gainontzeko esperimentuetan BPE erabili zelarik, *epoch* kopurua 30 izanik.

Taula honetan, 'eu-es' eta 'de-en' lerroek zentzu bakoitzeko aurre-entrenamendu sistemei erreferentzia egiten diete, eta hauen ondorengo lerroetan zentzu bakoitzean atzeranzko itzulpena egiteko dekodetze-teknika desberdinak erabiliz entrenatutako sistemei dagozkien balioak erakusten dira, etiketak gehituta edo ez. Sistema bakoitza identifikatzeko, 't.'-k etiketatutako corpusa erabiltzen duten sistemei erreferentzia egiten die, 'b.s.' *beam search*-en laburdura da, 'u.s.' *unrestricted sampling*-ena, eta 'r.s.' *restricted sampling*-ena.

Ikusten denez, 5.23. taulan erakutsitako balioekin alderatuta, nahiz eta atal honetan egindako esperimentuetarako sistema kopuru bikoitza baino gehiago entrenatu, CO_2 emisioen estimazioa ez zen asko handitu. Horren arrazoi nagusia oraingo honetan esperimentu bakoitzerako GPU bakarra erabiltzea izan zen; izan ere, esperimentu bat GPU anitzetan banatzerakoan denbora bat galtzen da GPUen artean komunikatzeko. Horregatik, memoria

Sistema	**Denbora (o)**	**Energia (kWh)**	**CO_2e (lbs)**
es-eu	81,93	32,36	30,88
eu-es	38,66	15,27	14,57
BT (b.s.)	71,90	28,40	27,10
BT (t.b.s.)	65,92	26,04	24,84
BT (u.s.)	75,66	29,89	28,51
BT (t.u.s.)	70,33	27,78	26,50
BT (r.s.)	70,83	27,98	26,69
BT (t.r.s.)	67,96	26,85	25,61
en-de	42,30	16,71	15,94
de-en	37,31	14,74	14,06
BT (b.s.)	51,53	20,35	19,42
BT (t.b.s.)	53,08	20,97	20,00
BT (u.s.)	54,37	21,48	20,49
BT (t.u.s.)	55,94	22,10	21,08
BT (r.s.)	52,26	20,64	19,69
BT (t.r.s.)	53,47	21,12	20,15
GUZTIRA			**355,53**

6.14 taula – Atzeranzko itzulpena egiteko dekodetze-teknika desberdinak erabilita entrenatutako sistemen entrenamendu-denbora, kontsumitutako energia eta estimatutako CO_2 emisioak.

nahikoa izanez gero, esperimentuak GPU bakarrean egitea eraginkorragoa da. Honetaz gain, 4.2. atalean aipatu moduan, kontuan hartu behar da aurrekoan erabilitako *OpenNMT*-rekin alderatuta, oraingoan erabilitako *Fairseq*-ek GPUen memoriaren erabilera baxuagoa egiten duela, entrenamenduaren momentu askotan memoria erabilera % 100 baino askoz txikiagoa izanik. Honen ondorioz, 6.14. taulako balioak interpretatzerakoan kontuan hartu behar da aurkeztutako balio hauek gain-estimazio bat direla. Etorkizunean, GPUen memoriaren erabilera eraginkorragoa egiteko moduak aztertuko dira, modu honetan CO_2 emisioei buruz egindako estimazioak zehatzagoak izango direlarik.

Atal hau amaitzeko, tesi honen helburutik haratago esperimentu multzo honetan egindako ekarpenak zerrendatuko ditugu:

4. Atzeranzko itzulpena egitean *tagged restricted sampling* teknika aplikatzea proposatu dugu, aurretik definitutako bi teknika konbinatuz, artearen egoera den *unrestricted sampling* dekodetze-teknikaren emaitza parekoak lortuz.

5. Itzulbideko corpus elebidunaren genero alborapena neurtu dugu, 'erizain' eta 'mediku' terminoak gazteleraz forma maskulino eta femeninoetan zenbat aldiz agertzen diren zenbatuz. Bestalde, sistemak entrenatzeko kontsumitutako energia eta sortutako CO_2 emisioen estimazioa egin dugu, antzeko esperimentuetarako erreferentzia bezala har daitekeena.

6.4 Azken sistemak: Itzulbideko corpus elebakarra datuen hautespenaren bidez gehitu eta giza ebaluazio sakona egitea

Helburua

Eskuragarri dauden corpus guztiekin azken sistemak entrenatzea, eu-es zentzuan atzeranzko itzulpena egiteko dekodetze-teknika hoberenak erabiliz; eta giza ebaluazio batean garatutako sistemak ebaluatzea.

Corpusak

Tesi honetan garatutako azken sistemak entrenatzeko eskura genituen corpus handienak gehitu genituen: Itzulbideko gaztelerazko corpus elebakarrak. Hauek orain arte erabilitako corpus elebidunak baino askoz handiagoak izanik, corpus elebidun eta elebakarraren arteko oreka mantentzeko datuen hautespena aplikatu genien Itzulbideko gaztelerazko corpus elebakarrei. Modu honetan, orain arte erabilitako 5M inguru esaldi pare elebidunei eta Galdakao-Usansolo Ospitaleko eta Basurtoko Unibertsitate Ospitaleko corpus elebakarretatik erauzitako beste 5M esaldi inguruei, Itzulbideko gaztelerazko corpus elebakarretik erauzitako beste 5M esaldi gehitu genizkien. Horrela, azken sistemetan erabilitako corpusean atzeranzko itzulpenaren bi-

dez erabilitako corpusa, corpus elebidunaren tamainaren bikoitza baino askoz handiagoa ez izatea bermatu zen, itzultzailearen kalitatea arriskuan jarri gabe (Poncelas *et al.*, 2018b). Laburpen moduan, 6.15. taulak atal honetan deskribatutako sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak aurkezten ditu.

Corpusa	**Esaldiak**
Domeinuz kanpoko 1-10 corpusak	4.929.728
SNOMED CT 2.0 + GNS-10 2.0 + SNOMED CT / COVID-19 + Elhuyar / COVID-19 terminoak	926.778
Itzulbide 2.0 + Basurtoko Unibertsitate Ospitaleko saio klinikoak	29.346
Galdakao-Usansolo Ospitaleko alta-txostenak, Basurtoko Unibertsitate Ospitaleko alta-txostenak, Basurtoko Unibertsitate Ospitaleko txosten ebolutiboak eta Itzulbideko gaztelerazko corpus elebakarrari datuen hautespena erabiliz erauzitako corpusa	10.145.926

6.15 taula – Azken sistemak entrenatzeko erabilitako corpusak eta haien esaldi kopuruak

Esperimentuak

Itzulbideko gaztelerazko corpus elebakar desberdinetatik txosten ebolutiboetatik erauzitako 18.667.813 esaldi eta traumatologiako 1.010.420 esaldi ez errepikatu erabili ziren hauei datuen hautespena aplikatzeko. Esaldi hauek guztiak aurreko atalean atzeranzko itzulpena egiteko emaitza hoberenak lortu zituzten dekodetze-tekniken bidez (*tagged restricted sampling* eta *unrestricted sampling*) itzuli ziren euskarara, eta datuen hautespena aplikatu ondoren aurretik erabilitako corpusetara gehitu zitzaien eu-es zentzuan bi sistema entrenatzeko.

Hori baino lehen, Itzulbideko corpus elebidunean esaldiak ondo parekatuak zeudela errepasatu genuen, hainbat zuzenketa eginez. Eguneratze hau es-eu eta eu-es zentzuetan entrenatutako azken sistemei aplikatu zitzaion, eta sistema hauek osasun-langileek ebaluatu beharrekoak izanda, emaitza hoberenak lortzen zituen BPE-dropout teknika erabili zen hitzen segmentaziorako.

Emaitzak

Lehenik eta behin, 6.16. taulak es-eu zentzuan corpus elebidun guztiak eta eguneratuenak erabiliz entrenatutako sistemaren IA emaitzak aurkezten ditu.

Noranzkoa	**BLEU↑**	**TER↓**	**METEOR↑**	**CHRF↑**
es-eu	33,18	49,92	46,38	60,45

6.16 taula – es-eu zentzuan corpus elebidun guztiak eta eguneratuenak erabiliz entrenatutako sistemaren IA emaitzak

Jarraian, 6.17. taulan eu-es zentzuan atzeranzko itzulpena egiteko emaitza hoberenak lortzen zituzten dekodetze-teknikak erabiliz garatutako azken sistemen IA emaitzak erakusten dira.

Sistema	**BLEU↑**	**TER↓**	**METEOR↑**	**CHRF↑**
tagged restricted sampling	54,35	32,62	69,43	72,13
unrestricted sampling	54,38	32,19	69,45	72,17

6.17 taula – eu-es zentzuan atzeranzko itzulpena egiteko emaitza hoberenak lortzen zituzten dekodetze-teknikak erabiliz garatutako azken sistemen IA emaitzak

Ondorioak

es-eu zentzuan, 6.16. taulan ikusten denez, aurretik 6.9. taulan aurkeztutako emaitzekin alderatuta emaitzak zertxobait baxuagoak dira oraingoan, baliteke BPE-dropout teknikak berezkoa duen ausazkotasunarengatik. Edonola ere, corpus eguneratuena erabiliz garatutako azken sistema erabili genuen corpus elebakarrak itzultzeko eta osasun-langileek ebaluatzeko.

eu-es zentzuan, 6.17. taulan ikusten dugunez, aurretik 6.10. taulan erakutsitako sistemen emaitzekin alderatuta, Itzulbideko gaztelerazko corpus elebakarra gehituta eta BPE-dropout aplikatuta emaitzak oraindik ere hobeak dira, eta atzeranzko itzulpena egiteko erabilitako bi metodo desberdinekin emaitza oso parekoak lortzen dira. Hortaz, Osakidetzan inplementatu beharreko eu-es sistema aukeratzeko, are beharrezkoagoa izango da osasun-langileek egin beharreko ebaluazioa.

Giza ebaluazioa

Oraingo honetan, sistemak ebaluatzeko sortutako itzulpenak post-editatzea eskatu zitzaien horretarako prest agertu ziren osasun-langileei. Ebaluatzaileak biltzeko, Itzulbideko corpus elebiduna biltzen lagundu zuten bolondresei mezu bat bidali zitzaien, eta horietatik 37k sistemak ebaluatzeko beren borondatea adierazi zuten.

Ebaluatutako sistemak 3 izan ziren: es-eu zentzuan garatutako azken sistema, eta eu-es zentzuan garatutako bi sistema hoberenak, zeintzuen IA emaitzak 6.16. eta 6.17. tauletan agertzen diren, hurrenez hurren. Sistema bakoitzarekin Itzulbideko corpus elebidunetik erauzitako proba multzoko esaldiak itzuli ziren, eta jatorrizko esaldiaren arabera errepikatutako esaldiak ezabatu ondoren, zentzu bakoitzean ausaz 500 esaldi erauzi ziren osasun-langileek ebaluatzeko. eu-es zentzuan, garatutako azken bi sistemek itzulpen berdina sortzerakoan esaldia behin bakarrik ebaluatu zen, eta zegokion emaitza bi sistemei egokitu zitzaien. Gainera, bi zentzuetan ebaluatu beharreko esaldietako bakoitza 2 ebaluatzaile desberdinek ebaluatu zuten, eta ebaluatzaile bakoitzari 100 esaldi ebaluatzeko eskatu zitzaion.

Ebaluazio hau egiteko PET[5] tresna erabili zen, eta honek itzulitako emaitzetatik post-edizio denbora eta HTER (Snover *et al.*, 2006) metrika erabili genituen sistemak ebaluatzeko. TER metrikaren antzera, HTER metrikaren balioa zenbat eta txikiagoa izan, hainbat eta handiagoa izango da itzultzaile automatikoaren kalitatearen estimazioa.

Hasteko, 6.18. taulak es-eu zentzuan garatutako azken sistemaren post-edizio denbora eta HTER balioen batez bestekoak erakusten ditu.

Noranzkoa	**Post-edizio denbora (s)**	**HTER**
es-eu	58,24	11,03

6.18 taula – es-eu zentzuan corpus elebidun guztiak eta eguneratuenak erabiliz entrenatutako sistemaren giza ebaluazioaren emaitzak

Azkenik, 6.19. taulan eu-es zentzuan atzeranzko itzulpenerako dekodetzeteknika desberdinak erabilita garatutako azken sistemen post-edizio denbora eta HTER balioen batez bestekoak aurkezten dira.

Ikusten denez, eu-es zentzuan garatutako sistemek es-eu zentzuan garatutako sistemak baino emaitza hobeak lortzen dituzte post-edizio denbora eta

[5] `https://github.com/wilkeraziz/PET`

Sistema	**Post-edizio denbora (s)**	**HTER**
tagged restricted sampling	39,87	5,87
unrestricted sampling	26,69	6,27

6.19 taula – eu-es zentzuan atzeranzko itzulpena egiteko emaitza hoberenak lortzen zituzten dekodetze-teknikak erabiliz garatutako azken sistemen giza ebaluazioaren emaitzak

HTERen arabera, metrika automatikoetan ikus litezkeen desberdintasunak konfirmatuz. eu-es zentzuan garatutako bi sistemei dagokienez, post-edizio denboraren arabera *unrestricted sampling* izango litzateke atzeranzko itzulpenerako dekodetze-teknika egokiena; eta HTERen arabera, *tagged restricted sampling*. Post-edizio denboran desberdintasunak handiak badira ere, kontuan hartu behar da PET tresnak erabiltzaileak esaldia bistaratzen duenetik bukatutzat ematen duen arteko denbora neurtzen duela. Tartean erabiltzaileak etenaldirik egiten badu, tresnak neurtutakoa errealitatetik desbideratu daiteke. Horrela, zuzentzeko 10 minutu baino gehiago behar izan dituzten esaldiak baztertzen baditugu, bi sistemen arteko aldeak asko murrizten dira, *tagged restricted sampling*-ekin 25,15 segundoko bataz bestekoarekin, eta *unrestricted sampling*-ekin 25,05 segundoko balioarekin. Desberdintasun hau oso txikia izanda, eta kontuan hartuta gure lehentasuna itzulpenen zuzentasuna dela, sistema hoberena aukeratzeko HTER metrikaren emaitzak lehenesten ditugu eta *tagged restricted sampling* sistema aukeratzen dugu Osakidetzan inplementatzeko.

7. KAPITULUA

Ondorioak, ekarpenak eta etorkizuneko lanak

Tesi honetan hainbat aldagai aztertu dira txosten klinikoak euskararen eta gazteleraren artean itzultzeko sistemak garatzerakoan. Kapitulu honen hasieran, helburu horretara hurbiltzen lagundu duten ondorioak zerrendatuko ditugu, 5. eta 6. kapituluetan agertu diren orden berean. Hobeto ulertzeko, ondorio bakoitza bere testuinguruan jarriko dugu, dagokion aldagaia eta horren inguruan egindako probak aipatuz.

Bestalde, tesi honen helburu nagusi den itzultzailea garatzetik haratago, kapitulu honen amaieran itzulpen automatikoaren arloan egindako ekarpenak aipatuko ditugu. Hauek ere tesiaren testuinguruan kokatzeko, gai horien inguruan egindako bestelako esperimentuekin batera aipatuko ditugu.

Azkenik, tesi honetan egindako lanen jarraipen moduan etorkizunean egin litezkeen lanak zerrendatuko ditugu.

7.1 Ondorioak

Tesiaren helburuetako bat terminologia klinikoa modu zuzenean itzultzea izanik, aurretik euskaratutako terminologia klinikoa eta tesian zehar COVID-19arekin lotuta sortutako terminologia berria bildu eta modu desberdinetan entrenamendu-corpusera gehitu da. Horrela, tesiaren hasieran itzultzaileak domeinu klinikora egokitzeko egindako esperimentuetan lehen ondorio honetara iritsi ginen:

1. Domeinuko testu elebidunik gabeko egoera batean, terminologia klini-

koa modu zuzenean entrenamendu-corpusera gehitzea lagungarria da.

Ordea, esperimentu multzo berean termino hauek esaldi artifizialetan txertatzea ez zela lagungarria ikusi genuen. Bestalde, aurrerago WMTko biomedikuntza atazan ingelesaren eta gazteleraren artean itzultzeko sistemak garatzerakoan ikusi genuen terminologia klinikoa gehitzeak ez zituela IA emaitzak hobetzen, eta sistemak sortutako itzulpenak laburragoak zirela.

Tesiaren beste helburuetako bat IA egiteko arkitektura desberdinak probatzea izan zen. Izan ere, tesia garatu den bitartean sare neuronalen arkitektura desberdinak izan dira momentu bakoitzean artearen egoera, eta horiek ere gure hizkuntza pare eta domeinurako egokienak zirela frogatu behar genuen. Horrela, arkitektura desberdinak probatu ondoren tesi honen bigarren ondoriora iritzi ginen:

2. Osasun-txostenak euskararen eta gazteleraren artean itzultzeko Transformer da arkitektura hoberena, baita atzeranzko itzulpena egiteko ere.

Bestalde, eskuragarri genituen domeinuko corpus gehienak elebakarrak izanik, tesian zehar egindako esperimentu asko atzeranzko itzulpenarekin lotutakoak izan dira. Horietarik lehenengoan, atzeranzko itzulpena egiteko sistema desberdinak probatu genituen, ohiko IAN sistemez gain, euskararako aurretik garatutako EOIA eta IAE sistemak ere probatuz. Honela, hurrengo ondoriora iritsi ginen:

3. Domeinuko corpus elebidunik gabeko egoera batean, EOIA aproposa izan daiteke atzeranzko itzulpena egiteko, IAEren emaitzak hobetuz.

Itzulpen sistema desberdinak garatzeaz gain, tesi honetan zehar corpus desberdinak bildu eta sistemetan integratzeko prestatu dira. Horietatik garrantzitsuena zalantzarik gabe Itzulbide proiektuan osasun-langileek bildutako corpus elebiduna izan da. Modu honetan, jarraian aurkezten dugun honako ondoriora iritsi ginen:

4. Itzulbideko corpus elebiduna oso baliagarria da txosten klinikoak euskararen eta gazteleraren artean itzultzeko, emaitzak nabarmen hobetuz.

Azkenik, corpusak aurreprozesatzeko modu desberdinak ere probatu dira. Horrela, tesian zehar aipatutako hizkuntz-identifikatzailea eta corpus garbiketa aplikatzeaz gain, hitzen segmentazioaren inguruan honako ondoriora iritsi ginen:

5. Txosten klinikoak euskararen eta gazteleraren artean itzultzeko, hitzen segmentaziorako BPE-dropout teknikak lortzen ditu emaitza hoberenak, corpusaren bi hizkuntzetan aplikatuz.

Hala ere, kontuan hartu behar da teknika honek BPEk baino askoz denbora gehiago behar duela; alde batetik, erregularizazio teknika bat izanda entrenamendua *epoch* gehiagotan egin behar delako, eta bestetik, *epoch* bakoitzarentzako entrenamendu-corpusaren kopia bat sortu behar delako. Hortaz, sistema desberdinak probatu behar izanez gero BPE arrunta erabiltzea gomendatzen da, BPE-dropout ebaluatu beharreko azkeneko sistemak garatzeko erabiliz.

7.2 Ekarpenak

Lehen ekarpena atzeranzko itzulpena egiteko erabilitako sistemekin erlazionatuta dago. Zentzu honetan, aurretik atzeranzko itzulpena egiteko IAN eta IAE sistemen irteerak konbinatzeko egindako lanari (Poncelas *et al.*, 2019), EOIA sistemak eta IAN arkitektura desberdinak batu dizkiogu. Honela, tesi honen lehenengo ekarpena honakoa litzateke:

1. Atzeranzko itzulpena egiteko 4 sistema desberdin erabilita sortutako corpusak batu ditugu, eu-es zentzuan sistema bakarra erabilita lortutako balioak hobetuz.

Bestalde, datuen hautespena aplikatzeko modu berri bat proposatu dugu, atzeranzko itzulpenarekin konbinatuta. Modu honetan, tesi honen bigarren ekarpenera iristen gara:

2. Lehen aldiz, atzeranzko itzulpenaren bidez sortutako corpusei datuen hautespena aplikatu diegu, hizkuntza pare batean hurbilpen guztien arabera 4 aldiz handiagoa den corpusa erabilita lortutako emaitzak hobetuz.

Datuen hautespena aplikatzeaz haratago, bere emaitzak atzeranzko itzulpena egiteko sistemen kalitatearen arabera birkalkulatzea proposatu dugu, ondorengo ekarpena eginez:

3. Datuen hautespenaren emaitzak atzeranzko itzulpena egiteko sistemen IA metriken eta hauen bidez sortutako corpusen aniztasun lexikalaren arabera birkalkulatzeko teknikak aztertu ditugu.

Hurrengo ekarpen nagusiak atzeranzko itzulpena egitean erabiltzen diren metodoekin erlazioa dauka:

4. Atzeranzko itzulpena egitean *tagged restricted sampling* teknika aplikatzea proposatu dugu, aurretik definitutako bi teknika konbinatuz, artearen egoera den *unrestricted sampling* dekodetze-teknikaren emaitza parekoak lortuz.

Azkenik, garatutako sistemek jendartean ekar litzaketen ondorioak aztertzeko asmoz, corpusen genero alborapena eta sistemak entrenatzerakoan ingurugiroan sortzen den eragina neurtu ditugu, honako ekarpena eginez:

5. Itzulbideko corpus elebidunaren genero alborapena neurtu dugu, 'erizain' eta 'mediku' terminoak gazteleraz forma maskulino eta femeninoetan zenbat aldiz agertzen diren zenbatuz. Bestalde, sistemak entrenatzeko kontsumitutako energia eta sortutako CO_2 emisioen estimazioa egin dugu, antzeko esperimentuetarako erreferentzia bezala har daitekeena.

7.3 Etorkizuneko lanak

- Erabiltzaileek itzultzaileen inguruan dituzten iritziak jaso eta, behar izanez gero, sistemak eguneratzea.
- Itzulbideko corpusean genero alborapenaren inguruan egindako azterketaren jarraipen moduan, garatutako sistemek genero estereotipoak ez erreproduzitzeko irtenbideak bilatzea, Osakidetzarekin elkarlanean. Norabide horretan, txosten klinikoak euskaratik gaztelerara itzultzerakoan genero alborapena neurtzeko proba multzo espezifiko bat definitzea.
- Garatutako itzultzaileak esaldi mailan entrenatuak izateak sistemen ebaluazioan ezartzen dituen mugak kontuan hartuta, txosten klinikoak euskararen eta gazteleraren artean dokumentu mailan itzultzeko sistemak garatzea.
- Garatutako sistemak idatzizko testura mugatuak izanda, hauek hizketara zein zeinu-hizkuntzara hedatzea, itzultzailearen aplikazio eremu posibleak handituz.

- Azkenik, garatutako itzultzaileak Osakidetzan inplementatzekoak izanik, euskararen eta gazteleraren artean itzultzeko diseinatuak izan direla kontuan izanda, Ipar Euskal Herriko osasun-langileekin elkarlanean txosten klinikoak euskararen eta frantsesaren artean itzultzeko sistemak garatzea.

Bibliografia

Agirre E., Alegria I., Arregi X., Artola X., Diaz de Ilarraza A., Maritxalar M., Sarasola K., eta Urkia M. XUXEN: A spelling checker/corrector for Basque based on two-level morphology. *Third Conference on Applied Natural Language Processing*, 119–125, Trento, Italy, March 1992. Association for Computational Linguistics. URL `https://aclanthology.org/A92-1016`.

ALPAC. Language and machines: Computers in translation and linguistics. a report by the automatic language processing advisory committee. *Nafional Academy of Sciences*, 1966.

Artetxe M., Labaka G., Agirre E., eta Cho K. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Bahdanau D., Cho K., eta Bengio Y. Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015. URL `http://arxiv.org/abs/1409.0473`. 15pp.

Banerjee S. eta Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72, Ann Arbor, Michigan, 2005. URL `https://www.aclweb.org/anthology/W05-0909`.

Barone A.V.M., Helcl J., Sennrich R., Haddow B., eta Birch A. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*, 2017.

Bawden R., DiÑunzio G.M., Grozea C., Jauregi Unanue I., Jimeno Yepes A., Mah N., Martinez D., Névéol A., Neves M., Oronoz M., Perez-de Viñaspre

O., Piccardi M., Roller R., Siu A., Thomas P., Vezzani F., VicenteÑavarro M., Wiemann D., eta Yeganova L. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. *Proceedings of the Fifth Conference on Machine Translation*, 660–687, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.76.

Biçici E. eta Yuret D. Optimizing instance selection for statistical machine translation with feature decay algorithms. *Transactions on Audio, Speech & Language Processing*, 23(2):339–350, 2015.

Bojar O., Chatterjee R., Federmann C., Haddow B., Huck M., Hokamp C., Koehn P., Logacheva V., Monz C., Negri M., Post M., Scarton C., Specia L., eta Turchi M. Findings of the 2015 workshop on statistical machine translation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-3001.

Britz D., Goldie A., Luong T., eta Le Q. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

Callison-Burch C., Osborne M., eta Koehn P. Re-evaluating the role of Bleu in machine translation research. *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.

Caswell I., Chelba C., eta Grangier D. Tagged back-translation. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 53–63, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-5206.

Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., eta Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar, 2014.

Chu C., Dabre R., eta Kurohashi S. An empirical comparison of domain adaptation methods for neural machine translation. *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 385–391, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-2061.

Chu C. eta Wang R. A survey of domain adaptation for neural machine translation. *Proceedings of the 27th International Conference on Computational Linguistics*, 1304–1319, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1111.

Currey A., Barone A.V.M., eta Heafield K. Copied monolingual data improves low-resource neural machine translation. *Proceedings of the Second Conference on Machine Translation*, 148–156, 2017.

Dew K.N., Turner A.M., Choi Y.K., Bosold A., eta Kirchhoff K. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, 85:56–67, 2018.

Dinu G., Mathur P., Federico M., eta Al-Onaizan Y. Training neural machine translation to apply terminology constraints. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3063–3068, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1294.

Edunov S., Ott M., Auli M., eta Grangier D. Understanding back-translation at scale. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1045.

Etchegoyhen T., Azpeitia A., eta Pérez N. Exploiting a large strongly comparable corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3523–3529, Portoroz, Slovenia, 2016.

Etchegoyhen T. eta Gete H. Handle with care: A case study in comparable corpora exploitation for neural machine translation. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3799–3807, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.469.

Etchegoyhen T., Martinez Garcia E., Azpeitia A., Labaka G., Alegria I., Cortes Etxabe I., Jauregi Carrera A., Ellakuria Santos I., Martin M., eta Calonge E. Neural machine translation of basque. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, 139–148. Alicante, Espainia, 2018.

Forcada M.L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J.A., Sánchez-Martínez F., Ramírez-Sánchez G., eta Tyers F.M. Apertium: A free/open-source platform for rule-based machine translation. *Neural Computation*, 25(2):127–144, 2011.

Graça M., Kim Y., Schamper J., Khadivi S., eta Ney H. Generalizing back-translation in neural machine translation. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 45–52, Florence, Italy, August 2019. Association for Computational Linguistics. URL `https://aclanthology.org/W19-5205`.

Haddow B., Birch A., eta Heafield K. *Machine translation in healthcare*. In The Routledge Handbook of Translation and Health. Routledge, 2021.

Heafield K. KenLM: Faster and Smaller Language Model Queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197, Edinburgh, UK, 2011. URL `http://mt-archive.info/WMT-2011-Heafield-2.pdf`.

Hochreiter S. eta Schmidhuber J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hokamp C. eta Liu Q. Lexically constrained decoding for sequence generation using grid beam search. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL `https://aclanthology.org/P17-1141`.

Hu J., Xia M., Neubig G., eta Carbonell J. Domain adaptation of neural machine translation by lexicon induction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2989–3001, Florence, Italy, July 2019. Association for Computational Linguistics. URL `https://aclanthology.org/P19-1286`.

Hutchins W.J. Machine translation: A brief history. *Concise history of the language sciences*, 431–445. Pergamon, 1995.

IHTSDO I.H.T.S.D.O. *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation, 2014.

Joanes Etxeberri Saria V. Edizioa. Donostia unibertsitate ospitaleko altatxostenak. *Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea*, 2014.

Johnson A.E., Pollard T.J., Shen L., Lehman L.w.H., Feng M., Ghassemi M., Moody B., Szolovits P., Anthony Celi L., eta Mark R.G. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016.

Kalchbrenner N. eta Blunsom P. Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709, Seattle, Washington, USA, 2013. URL `https://www.aclweb.org/anthology/D13-1176`.

Khayrallah H., Thompson B., Duh K., eta Koehn P. Regularized training objective for continued training for domain adaptation in neural machine translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 36–44, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL `https://aclanthology.org/W18-2705`.

Kingma D.P. eta Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klein G., Kim Y., Deng Y., Senellart J., eta Rush A.M. OpenNMT: Open-source toolkit for neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, 67–72, Vancouver, Canada, 2017. URL `http://arxiv.org/abs/1701.02810`.

Kocmi T., Federmann C., Grundkiewicz R., Junczys-Dowmunt M., Matsushita H., eta Menezes A. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*, 2021.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., eta Herbst E. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180, Prague, Czech Republic, 2007. URL `https://www.aclweb.org/anthology/P07-2045`.

Labaka G. *EUSMT: incorporating linguistic information to SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.* Doktoretza-tesia, UPV/EHU, 2010.

Labaka G., España Bonet C., Màrquez L., eta Sarasola K. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125, 2014.

Lample G., Denoyer L., eta Ranzato M. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Liu W. eta Cai S. Translating electronic health record notes from English to Spanish: A preliminary study. *Proceedings of BioNLP 15*, 134–140, Beijing, China, July 2015. Association for Computational Linguistics. URL `https://aclanthology.org/W15-3816`.

Magnini B., Altuna B., Lavelli A., Speranza M., eta Zanoli R. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. *CLiC-it*, 2020.

Marie B., Fujita A., eta Rubino R. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7297–7306, Online, August 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.acl-long.566`.

Mayor A. *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz.* Doktoretza-tesia, UPV/EHU, 2007.

McCarthy P.M. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity*. Doktoretza-tesia, University of Memphis, TN, 2005.

Och F.J. Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 160–167, Sapporo, Japan, 2003.

Och F.J. eta Ney H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Ott M., Edunov S., Baevski A., Fan A., Gross S., Ng N., Grangier D., eta Auli M. fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Papineni K., Roukos S., Ward T., eta Zhu W.J. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, AEB, 2002.

Perez-de Viñaspre O. *Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea*. Doktoretza-tesia, UPV/EHU, 2017.

Poncelas A., de Buy Wenniger G.M., eta Way A. Feature decay algorithms for neural machine translation. *21st Annual Conference of the European Association for Machine Translation*, 239–248, Alicante, Spain, 2018a.

Poncelas A., Popović M., Shterionov D., Maillette de Buy Wenniger G., eta Way A. Combining PBSMT and NMT back-translated data for efficient NMT. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 922–931, Varna, Bulgaria, September 2019. INCOMA Ltd. URL `https://aclanthology.org/R19-1107`.

Poncelas A., Shterionov D., Way A., de Buy Wenniger G.M., eta Passban P. Investigating backtranslation in neural machine translation. *21st Annual Conference of the European Association for Machine Translation*, 249–258, Alicante, Spain, 2018b.

Popović M. chrF: character n-gram f-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395, Lisbon, Portugal, September 2015. URL https://www.aclweb.org/anthology/W15-3049.

Provilkov I., Emelianenko D., eta Voita E. BPE-dropout: Simple and effective subword regularization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1882–1892, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.170.

Rei R., Stewart C., Farinha A.C., eta Lavie A. COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.213.

San Vicente I. eta Manterola I. PaCo2: A fully automated tool for gathering parallel corpora from the web. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1–6, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/231_Paper.pdf.

Sarasola I., Salaburu P., eta Landa J. *Hizkuntzen Arteko Corpusa (HAC)*. University of the Basque Country UPV/EHU (Euskara Institutua), Bilbao, Spain, 2015.

Sennrich R., Firat O., Cho K., Birch A., Haddow B., Hitschler J., Junczys-Dowmunt M., Läubli S., Barone A.V.M., Mokry J., eta Nădejde M. Nematus: a toolkit for neural machine translation. 2017.

Sennrich R., Haddow B., eta Birch A. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725, Berlin, Germany, 2015. URL http://arxiv.org/abs/1508.07909.

Sennrich R., Haddow B., eta Birch A. Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), 86–96, Berlin, Germany, 2016. URL https://www.aclweb.org/anthology/P16-1009.

Silva C.C., Liu C.H., Poncelas A., eta Way A. Extracting in-domain training corpora for neural machine translation using data selection methods. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 224–231, Brussels, Belgium, 2018.

Snover M., Dorr B., Schwartz R., Micciulla L., eta Makhoul J. A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 223–231. Cambridge, AEB, 2006.

Strubell E., Ganesh A., eta McCallum A. Energy and policy considerations for deep learning in nlp. *Computing Research Repository*, arXiv:1906.02243, 2019. URL http://arxiv.org/abs/1906.02243.

Sutskever I., Vinyals O., eta Le Q.V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104–3112, Montréal, Canada, 2014.

Templin M.C. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, Minneapolis, MN, 1975.

Tillmann C. eta Ney H. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133, 03 2003. ISSN 0891-2017. URL https://doi.org/10.1162/089120103321337458.

Vanmassenhove E., Shterionov D., eta Way A. Lost in translation: Loss and decay of linguistic richness in machine translation. *Proceedings of Machine Translation Summit XVII (Research Track)*, 222–232, Dublin, Ireland, 2019. URL https://www.aclweb.org/anthology/W19-6622.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., eta Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008, Long Beach, CA, 2017.

Wang L.L., Lo K., Chandrasekhar Y., Reas R., Yang J., Burdick D., Eide D., Funk K., Katsis Y., Kinney R.M., Li Y., Liu Z., Merrill W., Mooney

P., Murdick D.A., Rishi D., Sheehan J., Shen Z., Stilson B., Wade A.D., Wang K., Wang N.X.R., Wilhelm C., Xie B., Raymond D.M., Weld D.S., Etzioni O., eta Kohlmeier S. CORD-19: The COVID-19 open research dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.nlpcovid19-acl.1`.

Weaver W. Translation. memorandum. *Machine Translation of Languages: Fourteen Essays*. Reprinted in WN Locke and AD Booth, eds., 1949.

Yule G.U. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, UK, 1944.

Zeiler M.D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Zoph B., Yuret D., May J., eta Knight K. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.